PCT/JP03/04059

日 本 国 特 許 庁
JAPAN PATENT OFFICE

31.03.03

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年１２月 ４日

出願番号
Application Number: 特願２００２－３５２６４５

[ST.10/C]: [ＪＰ２００２－３５２６４５]

出願人
Applicant(s): 石原産業株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

２００３年 ５月 ９日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号 出証特２００３－３０３３９６３

【書類名】　特許願
【整理番号】　186699
【提出日】　平成14年12月 4日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/00
【発明者】
【住所又は居所】　滋賀県草津市西渋川二丁目３番１号　石原産業株式会社中央研究所内
【氏名】　石川　俊夫
【発明者】
【住所又は居所】　滋賀県草津市西渋川二丁目３番１号　石原産業株式会社中央研究所内
【氏名】　久米　隆志
【特許出願人】
【識別番号】　000000354
【住所又は居所】　大阪府大阪市西区江戸堀一丁目３番１５号
【氏名又は名称】　石原産業株式会社
【代理人】
【識別番号】　100062144
【弁理士】
【氏名又は名称】　青山　葆
【選任した代理人】
【識別番号】　100086405
【弁理士】
【氏名又は名称】　河宮　治
【選任した代理人】
【識別番号】　100098280
【弁理士】
【氏名又は名称】　石野　正弘

【先の出願に基づく優先権主張】

【出願番号】　特願2002-102743

【出願日】　平成14年 4月 4日

【手数料の表示】

【予納台帳番号】　013262

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【物件名】　委任状　1

【提出物件の特記事項】　手続補足書により提出する。

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　データ解析装置および方法

【特許請求の範囲】

【請求項１】　生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定するデータ解析装置であって、

生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力手段と、

(１)説明変数を選択する選択手段と、

(２)部分最小自乗法を実行して交差検証成績を計算する計算手段または前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算手段と、

(３)前記(２)の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを有し、

(４)前記（１）の選択手段と前記(２)の計算手段と前記(３)の評価判定手段とを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定手段とからなることを特徴とするデータ解析装置。

【請求項２】　目的変数が生体の状態であって、前記入力手段で入力するデータが生体の状態或いはそれを導出するデータであって、前記(２)の計算手段が部分最小自乗法を実行して交差検証成績を計算する計算手段であることを特徴とする請求項１に記載のデータ解析装置。

【請求項３】　目的変数が時間とともに確率的に発生する生体の状態の変化であって、前記入力手段で入力するデータが時間とともに確率的に発生する生体の状態の変化に関するデータであって、前記(２)の計算手段が前記生体の状態の

変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算手段であることを特徴とする請求項１に記載のデータ解析装置。

【請求項４】　さらに、前記の決定手段にて決定された部分最小自乗法モデルに採用されている説明変数を用い、統計的手法又は多変量解析手法によるモデルを構築する最終モデル決定手段を備えることを特徴とする請求項１、２又は３に記載のデータ解析装置。

【請求項５】　前記の選択手段において、説明変数を逐次取捨選択することを特徴とする請求項１～４のいずれかに記載のデータ解析装置。

【請求項６】　前記の選択手段において、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項１～４のいずれかに記載のデータ解析装置。

【請求項７】　前記の計算手段において、１個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１～６のいずれかに記載のデータ解析装置。

【請求項８】　前記の計算手段において、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１～６のいずれかに記載のデータ解析装置。

【請求項９】　前記計算手段において、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、交差検証成績の指標として当該誤差を用いることを特徴とする請求項７又は８に記載されたデータ解析装置。

【請求項１０】　前記関数が交差検証成績であることを特徴とする請求項１～９のいずれかに記載のデータ解析装置。

【請求項１１】　前記関数が交差検証成績と選択された説明変数の個数との関数であることを特徴とする請求項１～９のいずれかに記載のデータ解析装置。

【請求項１２】　前記の決定手段において、少なくとも交差検証成績を独立変数として持つ関数を改善しながら評価判定を繰り返すことを特徴とする請求項５に記載されたデータ解析装置。

【請求項１３】　前記（１）の選択手段と前記(２)の計算手段とを複数のコンピュータで実行させることを特徴とする請求項１～１２のいずれかに記載のデータ解析装置。

【請求項１４】　請求項１、２、３又は４で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力手段と、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定手段からなることを特徴とするデータ解析装置。

【請求項１５】　生体の状態を名義尺度、順序尺度或いは連続量で表現する目的変数とする請求項２に記載データ解析装置。

【請求項１６】　最終モデル決定手段が用いる前記の統計的手法又は多変量解析手法が、比例ハザード法又はパラメトリックな分布にあてはめた回帰分析法であることを特徴とする請求項２又は４に記載のデータ解析装置。

【請求項１７】　生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定するデータ解析方法であって、

生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力ステップと、

(１)説明変数を選択する選択ステップと、

(２)部分最小自乗法を実行して交差検証成績を計算する計算ステップまたは前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップと、

(３)前記(２)の計算ステップの結果を評価し、説明変数の採用、不採用を判定す

る評価判定ステップとを有し、

(４)前記（１）の選択ステップと前記(２)の計算ステップと前記(３)の評価判定ステップとを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなることを特徴とするデータ解析方法。

【請求項１８】　目的変数が生体の状態であって、前記入力ステップで入力するデータが生体の状態或いはそれを導出するデータであって、前記(２)の計算ステップが部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項１７に記載のデータ解析方法。

【請求項１９】　目的変数が時間とともに確率的に発生する生体の状態の変化であって、前記入力ステップで入力するデータが時間とともに確率的に発生する生体の状態の変化に関するデータであって、前記(２)の計算ステップが前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項１７に記載のデータ解析方法。

【請求項２０】　さらに、前記の決定ステップにて決定された部分最小自乗法モデルに採用されている説明変数を用い、統計的手法又は多変量解析手法によるモデルを構築する最終モデル決定ステップを備えることを特徴とする請求項１７、１８又は１９に記載のデータ解析方法。

【請求項２１】　前記の選択ステップにおいて、説明変数を逐次取捨選択することを特徴とする請求項１７～２０のいずれかに記載のデータ解析方法。

【請求項２２】　前記の選択ステップにおいて、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項１７～２０のいずれかに記載のデータ解析方法。

【請求項２３】　前記の計算ステップにおいて、１個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１７～２２のいずれかに記載のデータ解析方法。

【請求項２４】　前記の計算ステップにおいて、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１７～２２のいずれかに記載のデータ解析方法。

【請求項２５】　前記計算ステップにおいて、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、交差検証成績の指標として当該誤差を用いることを特徴とする請求項２３又は２４に記載されたデータ解析方法。

【請求項２６】　前記関数が交差検証成績であることを特徴とする請求項１７～２５のいずれかに記載のデータ解析方法。

【請求項２７】　前記関数が交差検証成績と選択された説明変数の個数との関数であることを特徴とする請求項１７～２５のいずれかに記載のデータ解析方法。

【請求項２８】　前記決定ステップにおいて、少なくとも交差検証成績を独立変数として持つ関数を改善しながら評価判定を繰り返すことを特徴とする請求項２１に記載されたデータ解析方法。

【請求項２９】　前記（１）の選択ステップと前記(２)の計算ステップとを複数のコンピュータで実行させることを特徴とする請求項１７～２８のいずれかに記載のデータ解析方法。

【請求項３０】　請求項１７、１８、１９又は２０で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなることを特徴とするデータ解析方法。

【請求項３１】　生体の状態を名義尺度、順序尺度或いは連続量で表現する目的変数とする請求項１８に記載データ解析方法。

【請求項３２】　前記の統計的手法又は多変量解析手法が、比例ハザード法又はパラメトリックな分布にあてはめた回帰分析法によるモデルを構築する最終モデル決定ステップとからなることを特徴とする請求項１８又は２０に記載のデータ解析方法。

【請求項３３】　生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定する、コンピュータにより実行されるデータ解析プログラムであって、

生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力ステップと、

(１)説明変数を選択する選択ステップと、

(２)部分最小自乗法を実行して交差検証成績を計算する計算ステップまたは前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップと、

(３)前記(２)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、

(４)前記（１）の選択ステップと前記(２)の計算ステップと前記(３)の評価判定ステップとを実行して部分最小自乗法モデルの少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなることを特徴とするデータ解析プログラム。

【請求項３４】　目的変数が生体の状態であって、前記入力ステップで入力するデータが生体の状態或いはそれを導出するデータであって、前記(２)の計算ステップが部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項３３に記載のデータ解析プログラム。

【請求項３５】　目的変数が時間とともに確率的に発生する生体の状態の変化であって、前記入力ステップで入力するデータが時間とともに確率的に発生する生体の状態の変化に関するデータであって、前記(２)の計算ステップが前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた

確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算ステップであることを特徴とする請求項３３に記載のデータ解析プログラム。

【請求項３６】　さらに、前記の決定ステップにて決定された部分最小自乗法モデルに採用されている説明変数を用い、統計的手法又は多変量解析手法によるモデルを構築する最終モデル決定ステップを備えることを特徴とする請求項３３、３４又は３５に記載のデータ解析プログラム。

【請求項３７】　前記の選択ステップにおいて、説明変数を逐次取捨選択することを特徴とする請求項３３～３６のいずれかに記載のデータ解析プログラム。

【請求項３８】　前記の選択ステップにおいて、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項３３～３６のいずれかに記載のデータ解析プログラム。

【請求項３９】　前記の計算ステップにおいて、１個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項３３～３８のいずれかに記載のデータ解析プログラム。

【請求項４０】　前記の計算ステップにおいて、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項３３～３８のいずれかに記載のデータ解析プログラム。

【請求項４１】　前記計算ステップにおいて、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、交差検証成績の指標として当該誤差を用いることを特徴とする請求項３９又は４０に記載されたデータ解析プログラム。

【請求項４２】　前記関数が交差検証成績であることを特徴とする請求項３３～４１のいずれかに記載のデータ解析プログラム。

【請求項４３】　前記関数が交差検証成績と選択された説明変数の個数との関数であることを特徴とする請求項３３～４１のいずれかに記載のデータ解析プログラム。

【請求項４４】　前記決定ステップにおいて、少なくとも交差検証成績を独立変数として持つ関数を改善しながら評価判定を繰り返すことを特徴とする請求項３７に記載されたデータ解析プログラム。

【請求項４５】　前記（１）の選択ステップと前記(２)の計算ステップとを複数のコンピュータで実行させることを特徴とする請求項３３～４４のいずれかに記載のデータ解析プログラム。

【請求項４６】　請求項３３、３４、３５又は３６で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなることを特徴とするデータ解析プログラム。

【請求項４７】　生体の状態を名義尺度、順序尺度或いは連続量で表現する目的変数とする請求項３４に記載データ解析プログラム。

【請求項４８】　前記の統計的手法又は多変量解析手法が、比例ハザード法又はパラメトリックな分布にあてはめた回帰分析法によるモデルを構築する最終モデル決定ステップとからなることを特徴とする請求項３４又は３６に記載のデータ解析プログラム。

【請求項４９】　上記の説明変数の選択において、初期状態では説明変数を全く含まないことを特徴とする請求項３７に記載されたプログラム。

【請求項５０】　上記の説明変数の選択において、初期状態では全説明変数を含むことを特徴とする請求項３７に記載されたプログラム。

【請求項５１】　上記の生体の状態が病気のタイプをあらわす測定値、病気の重篤度をあらわす測定値、病気のタイプをあらわす医療診断の結果、病気の重篤度をあらわす医療診断の結果、あるいはそれらを２次加工した数値であることを特徴とする請求項３７から５０のいずれかに記載されたプログラム。

【請求項５２】　請求項３３から請求項４８のいずれかに記載されたプログラムを記録した、コンピュータにより読み取り可能な記録媒体。

【請求項５３】　実質的にジーンバンクアクセッション番号がU15085、M23452、X52479、U70426、H57330及びS69790からなる遺伝子群の発現を検出すること

を特徴とするびまん性大細胞型Ｂリンパ腫の重篤度検定用の細胞内物質測定機材および測定方法並びにびまん性大細胞型Ｂリンパ腫の重篤度検定方法。

【請求項５４】　さらにジーンバンクアクセッション番号がU03398、M65066、AK001546、BC003536、X00437、U12979、H96306、AA830781及びAA804793からなる群から選択される少なくとも一つの遺伝子の発現を検出することを特徴とする請求項５３に記載のびまん性大細胞型Ｂリンパ腫の重篤度検定用の細胞内物質測定機材および測定方法並びにびまん性大細胞型Ｂリンパ腫の重篤度検定方法。

【請求項５５】　実質的にジーンバンクアクセッション番号がAA598572、AA703058及び AA453345からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の重篤度検定用の細胞内物質測定機材および測定方法並びに乳癌の重篤度検定方法。

【請求項５６】　さらにジーンバンクアクセッション番号がAA406242、 H73335、W84753、N71160、 AA054669、N32820及びR05667からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出することを特徴とする請求項５５に記載の乳癌の重篤度検定用の細胞内物質測定機材および測定方法並びに乳癌の重篤度検定方法。

【請求項５７】　実質的にジーンバンクアクセッション番号がW84753、H08581、AA045730及びAI250654からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

【請求項５８】　さらにジーンバンクアクセッション番号がAA448641、R78516、R05934、AA629838及びH53037からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出することを特徴とする請求項５７に記載の乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

【請求項５９】　実質的にジーンバンクアクセッション番号がAA434397、T83209、 N53427、N29639、 AA485739、AA425861、 H84871、T64312、 T59518及びAA037488からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方

法。

【請求項６０】　さらにジーンバンクアクセッション番号がAA406231の遺伝子産物を含む細胞内物質を検出することを特徴とする請求項５９に記載の乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

【請求項６１】　実質的にジーンバンクアクセッション番号がH11482、T64312及びAA045340からなる遺伝子産物を含む細胞内物質を検出することを特徴とする乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、生体の状態と遺伝子発現の量および／または細胞内物質の量との多変量解析処理並びにそれを基に可能となる測定機材、検定方法などに関するものである。

【０００２】

【従来の技術】

２０００年６月のヒトゲノムの解読宣言以降、ゲノムに書かれた遺伝情報がどのように発現して機能しているかのを解明するポストゲノム時代に突入したと言われている。ヒトゲノム計画の進展の中で、ゲノム発現状態を測定する方法論も進展してきた。トランスクリプトーム（ｍＲＮＡ）測定手段としてオリゴヌクレオチドアレイやマイクロチップが知られている。またプロテオーム（蛋白質）測定手段として、以前からある２次元電気泳動に加えて、最近では質量分析の方法が進歩してきた。また抗体チップなどの先進の技術も注目されている。これらの測定技術は、生体の状態パラメータを短時間に一挙に測定できることがそれまでの技術と比較して画期的であるといえる。

【０００３】

遺伝子発現状態を効率的に測定する技術として次のものがあげられる。トランスクリプトーム（ｍＲＮＡの総体）を特定するものとして、基盤に複数種のＤＮ

Ａを担持し、それに相補的なｍＲＮＡを検出するＤＮＡチップが知られている。代表的なＤＮＡチップには、遺伝子チップやＤＮＡマイクロアレイがある。また、プロテオーム（蛋白質の総体）を特定するものには、２次元電気泳動、抗体チップ、質量スペクトルを用いるものがある。またメタボローム(代謝中間体を含めた代謝産物の総体)を測定する手法も質量分析などによって試みられており、進展が見られる。

【０００４】

生体内の細胞の状態は遺伝子産物の発現によってよく記述されるため、従来の診断マーカーでは情報が不足している場面でも、精度のより高い診断が可能になるという期待も出てきている。たとえば、次のような研究があげられる。

【０００５】

P. O. Brownらは、ＤＮＡチップによってリンパ腫患者の細胞のトランスクリプトームを測定し、クラスター解析によって悪性と良性のリンパ腫（ＤＬＢＣＬ）を別クラスターに分離した（Nature 403(3), 503-11 (2000)）。しかし、これは因果関係（相関関係）のモデルを得る方法ではなく、どの遺伝子がどの程度重要かを判断できない。

【０００６】

A. Alaiyaらは、２次元電気泳動によって子宮がん患者４０人の細胞のプロテオームを測定し、うち２２人のデータから部分最小自乗法診断モデルを構築し、悪性度を説明した（Int. J. Cancer, 86, 731-36 (2000); Electrophoresis, 21, 1210-17 (2000); 国際公開 WO 00/70340)。その際、全変数モデルにおいて１５５３変数からloadingの大きな１７０変数に限定することによって交差検証成績がよくなり（$Q^2=0.84$）、残り１８患者の深刻度（３段階）を１１／１８の比率で正答した。交差検証法がモデル構築の際の指標になるという考えが表明されている。しかし、この方法では、loadingを得る際にまず全変数モデルが成立しなければならない。また、それ以外の変数選択手法が考案されていない。

【０００７】

J. Khanらは、ＤＮＡチップによって小児がん患者の細胞を測定し、ニューラルネットワークによって悪性度を説明した（Nature Medicine, 7(6), 673-79 (2

001))。小児がん（SRBCT）患者８８人のトランスクリプトーム（６５６７遺伝子）を測定し、うち６３人のデータから主成分分析によって１０次元に圧縮し、次に、人工ニューラルネットワーク診断モデルを構築した。ここで、影響力のある上位遺伝子を交差検証法によって絞り込み、９６遺伝子で最良の成績(100%)を得た。このモデルで残り２５人を予測し、９３～１００％の結果を得た。しかし、この方法でも、影響力を得る際にまず全変数モデルが成立しなければならない。またそれ以外の変数選択手法が考案されていない。１０次元のような少ない変数の場合を扱えるが、変数の数が膨大な場合には適用できない。

【０００８】

【特許文献１】

国際公開公報 WO 00/70340

【非特許文献１】

P.O. Brownら、Nature 403(3), 503-11 (2000)

【非特許文献２】

A. Alaiyaら、Int. J. Cancer, 86, 731-36 (2000)

【非特許文献３】

A. Alaiyaら、Electrophoresis, 21, 1210-17 (2000)

【非特許文献４】

J. Khanら、Nature Medicine, 7(6), 673-79 (2001)

【非特許文献５】

P. Gramaticsら, Chemosphere 38(5), 1371-78 (1999)

【０００９】

【発明が解決しようとする課題】

従来の診断マーカーでは情報が不足している場面でも、遺伝子発現情報を活用することで、より精度（解像度）の高い診断が可能になるという期待も出てきている。遺伝子発現状態の測定結果は、膨大な情報量が得られることが従来にはなかった特徴であり、逆に情報量が多いために、効果的なデータ処理なくしてデータの活用はありえない。したがって、有用な知識を獲得するためには効果的な情報処理が欠かせない。前に説明したように、現状ではクラスター解析を中心とす

る方法が用いられているが、主成分分析などの方法も採用されている。クラスター解析や主成分分析は、教師付学習方法ではないため、病状の因果関係（相関関係）のモデルを得ることはできない。すなわち、どの遺伝子がどの程度重要かを解析結果から得ることができないのが難点である。一方、部分最小自乗法は次元圧縮とモデルフィットを同時に行なう強力な多変量解析手法であるが、変数の数が膨大になった場合にしばしば有意な結果が得られない事態に直面する。したがって、膨大な遺伝子発現情報などから有用な知識を獲得できるような効果的な情報処理が望まれている。また、そのような情報処理の結果を基にした効率的な測定機材、検定処理などが期待されている。

【００１０】

この発明の目的は、多変量の遺伝子発現情報、細胞内物質情報の効果的な情報処理を提供することである。

また、この発明の目的は、効率的な検定処理を提供することである。

【００１１】

【課題を解決するための手段】

本発明に係るデータ解析装置は、生体の状態または時間とともに確率的に発生する生体の状態の変化を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とする相関モデルを決定するデータ解析装置であって、生体の状態或いはそれを導出するデータまたは時間とともに確率的に発生する生体の状態の変化に関するデータと、複数の遺伝子発現の量および／または細胞内物質の量からなるサンプルの集合を入力する入力手段と、（１）説明変数を選択する選択手段と、（２）部分最小自乗法を実行して交差検証成績を計算する計算手段または前記生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算して得られた確率を、仮定した分布に基づいた変換または仮定を前提としない変換をし、該変換結果を目的変数とする部分最小自乗法を実行して交差検証成績を計算する計算手段と、（３）前記（２）の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを有し、（４）前記（１）の選択手段と前記（２）の計算手段と前記（３）の評価判定手段とを実行して部分最小自乗法モデルの

少なくとも交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定手段とからなる。前記選択手段は、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算手段は、たとえば、１個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定手段は、たとえば、計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。あるいは交差検証成績ではなく、少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数が改善するかどうかを評価判定の基準として用いることもできる。決定手段は、たとえば、選択手段と計算手段と評価判定手段とを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択手段と計算手段とを複数のコンピュータで実行させることもできる。こうして、相関モデルを構成するとき、交差検証成績を基準に最適化させることにより説明変数を取捨選択し、説明変数の数を減らす次元圧縮をして良好なモデルを得る。

【００１２】

上述の、仮定した分布に基づいた変換または仮定を前提としない変換は、生体の状態の変化の確率が説明変数の多項式で解析できるようにするために行なうものである。分布を仮定した場合には、確率を対数変換後に負の数にしたものを状態の変化を観測した時間で割るという変換、確率を対数変換後に負の数にしたものをさらに対数にしたものを状態の変化を観測した時間で割るという変換、または確率を1より減じたものをプロビット変換したものを計算しを状態の変化を観測した時間で割るという変換などがが考えられる。一方、分布を仮定しない場合にはロジット変換といった方法が考えられる。変換の方法は分布にどのような仮定が成り立つかどうかあるいはなりたたないかどうかを判断することにより、それぞれの場合に応じて適切に選ぶことができる。少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数としては、たとえば、前記誤差の代

表値と選抜された説明変数の数の関数が考えられ、あるいはその他の独立変数を含むものであってよい。望ましくは、関数は誤差の代表値の単調減少関数であり、説明変数の数の単調減少関数である。計算量を増やさないためには簡単に計算できる関数が望ましい。具体的には-PRESS×alpha^{NP}という関数が考えられる。ここでPRESSは予測残差自乗和であり、ＮＰは採用された説明変数の数であり、alphaは１または１より大きい実数である。また、-PRESS×(NP+beta)^{gamma}や-PRESS×(beta-NP)$^{-gamma}$なる関数も考えられる。ここで、gammaは正の実数である。

【００１３】

説明変数の個数を少なくすると、通常の統計的手法または多変量解析手法が適用可能になる。本発明では、部分最小自乗法を用いて得られた説明変数を統計的手法又は多変量解析手法の説明変数として、より良好なモデルを得る。統計的手法又は多変量解析手法としては、重回帰分析法、線型判別分析法、適応最小自乗法、ロジスティック回帰分析法、比例ハザード解析法、マハラビノス距離を用いる判別分析法、ｋＮＮ法、人工ニューラルネットワークなどが挙げられる。

【００１４】

本発明者等は、また、Q^2やPRESS値などの交差検証成績に加えて、説明変数の個数を第２の独立変数として含む関数を最適化することで選抜される説明変数を任意に絞り込むことができることを新たに見出した。通常の統計的手法や多変量解析手法では、抽出される説明変数の個数ＮＰの望ましい範囲がサンプル数との兼ね合いで決まっている場合がある。そのような場合、関数を、目的とする選抜数によって任意に変更できる。関数形をたとえば-PRESS×alpha^{NP}とした場合、説明変数の個数を数個から数十個に絞り込むためには通常は定数alphaとして１．０～３．０の値が望ましい。より望ましくは、alphaは１．０～２．０の値となる。他の関数形f(PRESS，NP)であっても、実際に選択される説明変数の数MPおよびその時のPRESS値PRESS_MPの周辺で、f(PRESS_MP÷alpha，MP+1)≒f(PRESS_MP，MP)となるような関数は、変数選択という点では同様の効果を持つ場合がある。こうして、適当な関数形を用いることにより、望ましい範囲の個数ＮＰの説明変数を選抜できる。このようにして、交差検証成績を用いて決定されたモデルに採

用されている説明変数をさらに絞り込むと、統計的手法又は多変量解析手法によるモデルを構築できる。したがって、その性質が十分解明されている統計的手法又は多変量解析手法を採用して解析を加えることができる。

【００１５】

また、目的変数として、時間とともに確率的に発生する生体の状態の変化から導出された量を用いて、時間とともに確率的に発生する生体の状態の変化と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定できる。「時間とともに確率的に発生する生体の状態の変化」とはたとえば生存時間である。ここで、前述の部分最小自乗法に、カプラン・マイヤー法又はカトラー・エデラー法と、ロジット(logit)変換とを組み合わせる。部分最小自乗法での目的変数は、時間とともに確率的に発生する生体の状態の変化に関するデータにカプラン・マイヤー法又はカトラー・エデラー法による生命表を適用して変化の発生しなかったものの確率を計算し、これをロジット変換した値である。ロジット(logit)値とは、分類分けされたデータの、ある分類の割合（確率）Ｐを基に、次式logit＝log{P/(1-P)}にて計算される値である。ロジット値を目的変数とする部分最小自乗法を実行して交差検証成績を計算する。こうして、先に説明したのと同様に、部分最小自乗法の交差検証成績を考慮した説明変数の抽出を行って、生存時間解析を行える。

【００１６】

説明変数の個数を少なくすると、通常の統計的手法または多変量解析手法が適用可能になる。そこで、決定されたモデルに採用されている説明変数を用い、時間とともに確率的に発生する生体の状態の変化を説明する統計的手法又は多変量解析手法によるモデルを構築する。たとえば、ロジット値を目的変数として求めた説明変数を用いて、他の統計的手法又は多変量解析手法（たとえば比例ハザード法や、パラメトリックな分布にあてはめた回帰分析法）を行なうことによって、より良好なモデルを得ることができる。比例ハザード法とは、Coxによって考案された方法であり、生存率の解析に時間を考慮し、かつ、多変量を扱える。比例ハザード法では、観測されている個々ごとにハザード値と呼ばれる生存率を左右する値があり、それを導く関数がある（モデルが仮定されている）として解析

される。カプランーマイヤー法は、集団全体または群ごとの生存率の推移を示す。また、パラメトリックな分布とは、ガウスが提案した正規分布から計算された確率分布のことであり、生存時間解析では指数分布、ワイブル分析、対数正規分布が用いられる。指数分布などへの当て嵌めで、数式中に多項式があり、前述の部分最小自乗法の交差検証成績を考慮した説明変数の抽出が適用される。

【００１７】

入力手段で説明変数として入力される複数の遺伝子の発現量および／または細胞内物質の量とは、必ずしも物質の絶対的な濃度の測定値に限定されるものではなく、加工計算された値、相対的な値、間接的に物質量を表す量などでもよい。たとえば、質量スペクトルで蛋白質の発現量を測定することができることを応用して、生体の状態を表わす目的変数と、質量スペクトルとを直接関係づける相関モデルを構築することができる。またAffymetrix社タイプのＤＮＡチップ(ジーンチップ)では、単一のスポットが単一の遺伝子発現を特定するとは限らず、複数個のスポットが集まってはじめて単一の遺伝子発現を特定することもある。ここでもまた、各スポットの測定量を説明変数として、直接、生体の状態を説明する相関モデルを得ることができる。更には、タンパク質の電気泳動パターンの各ピークは単一のタンパク質に帰属できず、複数個のタンパク質の重ねあわせであることも多い。このような場合にも生体の状態を説明する説明変数として各ピーク強度を用いることができる。このことは、上述のAlaiyaらは子宮癌の診断の説明変数として電気泳動パターンのピーク強度を採用していることから明らかである。前述のようにポストシークエンス時代のトランスクリプトーム解析、プロテオーム解析、メタボローム解析という研究分野では、生体(細胞)内の物質を総体として把握することから出発することを特徴とする実験的アプローチが注目されている。ひとつひとつの物質の絶対的定量は必須事項ではなく、これらの実験方法によって定量される物質の量を直接、間接に表現する測定値やその加工計算値が、生体の状態を説明する説明変数と成り得る。また以上の物質量を表現する説明変数以外に、場合によっては問診データなどの他の説明変数を追加すると、さらに有効な解析結果が得られる場合もある。

【００１８】

本発明に係るデータ解析方法は、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析方法であって、生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力ステップと、(１)説明変数を選択する選択ステップと、(２)部分最小自乗法を実行して交差検証成績を計算する計算ステップと、(３)前記(２)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、(４)前記（１）の選択ステップと前記(２)の計算ステップと前記(３)の評価判定ステップとを実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなる。前記選択ステップは、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算ステップは、たとえば、１個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定ステップは、たとえば、計算ステップの結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定ステップは、たとえば、選択ステップと計算ステップと評価判定ステップとを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択ステップと計算ステップとを複数のコンピュータで実行させることもできる。

【００１９】

本発明に係るデータ解析プログラムは、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定する、コンピュータにより実行されるデータ解析プログラムであって、生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力ステップと、(１)説明変数を選択する選択ステップと、(２)部分最小自乗法を実行して交差検証成績を計算する計算ステップと、(３)前記(２)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステッ

プとを有し、（４）前記（１）の選択ステップと前記（２）の計算ステップと前記（３）の評価判定ステップとを実行して少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなる。

【００２０】

上記のデータ解析プログラムにおいて、前記選択ステップは、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算ステップは、たとえば、１個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定ステップは、たとえば、計算ステップの結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、少なくとも当該誤差の代表値を独立変数として持つ関数である当該誤差の誤差の代表値の単調減少関数の値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定ステップは、たとえば、選択ステップと計算ステップと評価判定ステップとを繰り返し実行して少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数を改善し続けて部分最小自乗法モデルを決定する。また、選択ステップと計算ステップとを複数のコンピュータで実行させることもできる。さらには、上記の説明変数の選択において、たとえば、初期状態では説明変数を全く含まないか、或いは、初期状態では全説明変数を含むこともできる。

【００２１】

上記のデータ解析プログラムにおいて、上記の生体の状態は、たとえば病気のタイプあらわす測定値、病気の重篤度をあらわす測定値、病気のタイプをあらわす医療診断の結果、病気の重篤度をあらわす医療診断の結果、あるいはそれらを２次加工した数値である。例えば後の実施例で示すように、患者の生存時間を予測することは、QOLを含めた治療計画や人生設計などを判断する上で重要な情報をもたらすものであり、社会的に価値のある診断モデルを提供することができる。また癌の再発可能性を予測することは、QOLを考慮した治療計画を立案し、医

師または当の患者が選択の判断をするうえで、貴重な情報をもたらすものである。

【００２２】

また、本発明は、前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力手段と、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定手段からなるデータ解析装置、前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなるデータ解析方法及び前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなるデータ解析プログラムも包含する。

【００２３】

本発明に係るコンピュータにより読取可能な記録媒体は、上記のいずれかのプログラムを記録する。

【００２４】

本発明に係るびまん性大細胞型Ｂリンパ腫の重篤度検定用の細胞内物質測定機材および測定方法並びにびまん性大細胞型Ｂリンパ腫の重篤度検定方法は、実質的にジーンバンクアクセッション番号がU15085、M23452、X52479、U70426、H57330及びS69790からなる遺伝子群の発現を検出する。さらに、ジーンバンクアクセッション番号がU03398、M65066、AK001546、BC003536、X00437、U12979、H96306、AA830781及びAA804793からなる群から選択される少なくとも一つの遺伝子の発現を検出してもよい。

【００２５】

また、本発明に係る乳癌の重篤度検定用の細胞内物質測定機材および測定方法並びに乳癌の重篤度検定方法は、実質的にジーンバンクアクセッション番号がAA598572、AA703058及び AA453345からなる遺伝子産物を含む細胞内物質を検出する。さらに、ジーンバンクアクセッション番号がAA406242、 H73335、W84753、N

71160、 AA054669、N32820及びR05667からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出してもよい。

【００２６】

また、本発明に係る乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法は、実質的にジーンバンクアクセッション番号がW84753、H08581、AA045730及びAI250654からなる遺伝子産物を含む細胞内物質を検出する。さらに、ジーンバンクアクセッション番号がAA448641、R78516、R05934、AA629838及びH53037からなる群から選択される少なくとも一つの遺伝子産物を含む細胞内物質を検出してもよい。

【００２７】

また、本発明に係る乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法は、実質的にジーンバンクアクセッション番号がAA434397、T83209、 N53427、N29639、 AA485739、AA425861、 H84871、T64312、T59518及びAA037488からなる遺伝子産物を含む細胞内物質を検出する。さらに、ジーンバンクアクセッション番号がAA406231の遺伝子産物を含む細胞内物質を検出してもよい。

【００２８】

また、本発明に係る乳癌の再発性検定用の細胞内物質測定機材および測定方法並びに乳癌の再発性検定方法は、実質的にジーンバンクアクセッション番号がH11482、 T64312及びAA045340からなる遺伝子産物を含む細胞内物質を検出する。

【００２９】

細胞内物質測定機材としては、ＤＮＡマイクロアレイ、ジーンチップ、オリゴＤＮＡ型のＤＮＡチップ、電気化学ＤＮＡチップ(ＥＣＡチップ)、繊維型ＤＮＡチップ、磁性ビーズＤＮＡチップ(PSS)、糸巻きＤＮＡチップ(PSS)、などのＤＮＡチップ、マクロアレイ、抗体チップ、測定用試薬キットなどが挙げられる。また、上記の機材を適宜組み込んだ測定機械であってもよい。

【００３０】

【発明の実施の形態】

以下、添付の図面を参照して本発明の実施の形態を説明する。

以下に、選択された生体の状態と遺伝子発現の量および／または細胞内物質の量との相関モデルの決定について説明する。ここで、遺伝子発現の用語は、ｍＲＮＡ発現(トランスクリプトーム)や、ｍＲＮＡによる翻訳の結果として生じる蛋白質(プロテオーム)を含むものとして用いる。また、細胞内物質の量とはここではたとえば、代謝中間体を含めた代謝産物全部であるメタボロームを意味する。たとえば、トランスクリプトーム(ｍＲＮＡ)やプロテオーム(蛋白質)の解析において、各サンプルデータは、生体の状態と遺伝子発現の量などからなる。各サンプルはたとえば１０００個以上の膨大な遺伝子発現の量を含む。生体の状態は、たとえば病気のタイプまたは病気の診断指標であるが、より一般的には生体情報であればよい。「病気の診断指標」には、病気の進行度合いのほか、病気のタイプ、重篤度、深刻度などの表現で表わされるものも含む。ここで、遺伝子発現の量などの測定データは膨大な情報量からなるので、コンピュータを用いた効率的な多変量解析が必要である。

【００３１】

データ収集において、予めいくつかのサンプルについて生体の状態（たとえば診断指標）を判定し、また、そのサンプルされたものから細胞液を獲得し、その細胞液中の多くの遺伝子産物の発現の量などを測定する。本発明の実施の形態のデータ解析では、こうして得られた遺伝子産物の発現の量などと生体の状態（たとえば診断指標）を入力し、相関モデル（たとえば部分最小自乗法モデル）を得る。ここで、コンピュータによる多変量解析プログラムを用いて、診断指標を目的変数とし、遺伝子発現の量および／または細胞内物質の量を説明変数とする因果関係型の解析を行なって、各説明変数の重要性や影響度に関する情報を得る。また、前記目的変数は、必ずしも測定値そのものである必要はなく、ロジット変換を行なった値や群を表す離散値を用いても良く、その場合、より有意な解析結果を得ることもできる。

【００３２】

本発明者らは、遺伝子発現による医療診断という分野において、データ解析における交差検証（cross validation）の成績を少なくとも独立変数のひとつとして持つ関数を最適化するように変数を選択することによって良好な相関モデル（

たとえば部分最小自乗法モデル）が得られることを見出した。交差検証法では、手持ちのデータを複数群に分割し、その一部のデータ群（訓練集合）だけを使ってフィットしたモデルを用いて残る別のデータ群（テスト集合）を予測することによって、モデルの予測力を試す。通常の部分最小自乗法（PLS）においては潜在変数の次元選択に交差検証法が用いられているが、ここでは、部分最小自乗法において、潜在変数を１次元に固定し、１以上の入力変数（説明変数）を逐次取捨選択しながら、交差検証成績（たとえば平方和の予測誤差）を少なくとも独立変数のひとつとして持つ関数を最適化した。その結果、全変数を採用した場合には有意な相関モデルを得られなかった場合にも、良好でかつ予測力のある相関モデルが得られることが判明したのである。この交差検証法を用いた変数選択の逐次取捨選択により、安定な相関モデルが得られる。また本発明者らは、関数形を適切に設定することによって説明変数を絞り込むことにより、部分最小自乗法以外の統計学又は多変量解析の良好な相関モデルを得ることが可能となり、それぞれ生体の状態を記述する目的変数にふさわしい相関モデルを得ることができることを見出した。なお、ここでいう「最適化」とは、交差検証成績が、説明変数を取捨選択するための、そのときの解析条件の範囲で、改善がみられなくなるまで改良したことを意味しており、交差検証成績がすべての説明変数の組合せの中で最適なものを見出したという意味ではない。この変数選択手法を用いると、病状を決定する因子を少数に特定し、廉価な診断用材料（ＤＮＡチップ、抗体チップ、ＤＮＡ含有ベクターなど）を設計でき、それ自体独自の価値を持つものである。また、この変数選択手法は、予め設定される各種の変数選択条件と共に運用することが可能である。

【００３３】

上に述べたように、説明変数は、交差検証成績を基準に逐次取捨選択される。ここで、取捨選択のため、交差検証成績を少なくとも独立変数のひとつとして持つ関数を用いる。説明変数を追加する場合は、その説明変数について、前記関数が改善されなかったと判定された場合には当該説明変数を除外し、改善されたと判定された場合には当該説明変数を追加する。また、説明変数を除外する場合は、その説明変数について、前記関数が改善されなかったと判定された場合には当

該説明変数を除外せず、改善されたと判定された場合には当該説明変数を除外する。ここで、１以上の説明変数を選択した場合に、交差検証成績評価は次のように進める。ｎ個のサンプルからいくつかのサンプルを逐次除外して部分最小自乗法モデルを求め、各モデルにおいて除外したサンプルの遺伝子発現の量から予測される生体の状態を示す目的変数と、除外したサンプルの生体の状態を示す目的変数との各々の誤差の代表値を求める。「代表値」とは、和、平均、最大値、中位値、最頻値などのデータを特徴づける値をいう。そして、当該誤差の代表値を少なくともひとつの独立変数とする関数が小さくなった場合に、交差検証成績が改善されたと判定し、当該説明変数を追加または削除する。この交差検証成績評価を、説明変数を取捨選択しながら逐次繰り返して、前記関数を改善し続ける。改善されなくなれば交差検証成績を最適化したとして説明変数の取捨選択を終了する。その結果、取捨選択により絞り込んだ数の説明変数からなる最適な部分最小自乗法モデルが得られる。具体的には、計算手段において計算される交差検証成績の数値指標として予想残差自乗和(PRESS)を採用し、評価判定手段において予想残差自乗和の値が説明変数あたり一定の閾値以下の比率で小さくなる場合に、その説明変数を採用すると判定することにより、上記の処理は実行可能である。

【００３４】

因果関係型の解析手法においてはオーバーフィット（over fitting）を避けるための工夫が必要となる。ここでいうオーバーフィットとは、説明変数が多すぎるためにたまたま予測結果と実績とが一致するものの、本当の相関関係をとらえ損なっているため、モデルフィットに用いたデータ以外に予測能力を持たないことをいう。ここでは、相関モデルとして部分最小自乗法を用いるが、部分最小自乗法は次元圧縮とモデルフィットを同時に行なう強力な多変量解析手法であり、オーバーフィットの問題に比較的強いとされている。しかし遺伝子発現状態解析のように膨大な変数を扱う場合には、有意な結果が得られない事態に直面する。従来技術として説明したAlaiyaやKhanの手法は全変数モデルが有意に成立することを前提としているので、変数の絞込みには一般的には適用できない。これに対し、本発明では、交差検証予測結果を最適にするように変数を絞り込むことによ

り、オーバーフィットを減らすことができた。また、本発明は、前記Khanの手法とは異なり、主成分分析などの前処理を介さない方法である。従来技術では、説明変数が膨大な場合には、有意なモデルを得ることができないことから、予め、全説明変数を基にたとえば、主成分分析などで次元圧縮する前処理をし、これによって得られた説明変数によって解析する方法が用いられる。しかし、この方法では、構成したモデルで予測を行なうためには、モデル構成の基となった全説明変数が必ず必要となり、たとえば、説明変数が遺伝子発現の量であれば、診断用遺伝子チップに担持する遺伝子としては、モデル構成に用いた遺伝子の全てが必要となるか、または別の手法を用いて変数選択することが必要となる。一方、本発明においては、説明変数の選択によって説明変数を絞り込んでいるので、たとえば、説明変数が遺伝子発現の量であれば、診断用遺伝子チップに担持する遺伝子は、選択された説明変数に相当する遺伝子を担持すれば良いことになる。

【００３５】

なお、Todeschiniらは、有機化合物の大気中の分解を予測するため、遺伝的アルゴリズムによって交差検証成績を最適化するように変数選択を行ない、重回帰モデルを得ている（P. Gramatics, V. Consonni & R. Todeschini, Chemosphere 38(5), 1371-78 (1999)）。５３化合物と１７５記述子でモデル構築を行ない（Q^2＝0.79)、７変数が選択され、９８化合物の予測を行なった（Q^2＝0.75)。交差検証成績を最適化するように変数選択を行なっている点では、本実施形態と同様の手法である。しかし、重回帰モデルを採用しているために、説明変数の選択過程を通じて選択される変数は少数個にとどまらざるを得ず、複数の遺伝子発現の量および／または細胞内物質の量の解析には適用できない。本発明者らの調査した範囲では、Q^2やPRESS値を最適化する方法では、選抜される説明変数は百程度から数百程度にわたり、重回帰モデルでは解析が不能となる。またTodeschiniらは、説明変数を絞り込むための有効な方法について言及していない。これは、もともとの説明変数の候補がたかだか１７５個であり、説明変数を絞り込むために特別の工夫をする必要がないからである。遺伝子発現解析の分野はこれとは全く異なり、数十から数百のサンプル数に対して、数百から数千、数万の説明変数候補が存在する。したがってこれまでとは異なる工夫が必要となる。

【００３６】

本実施形態では、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するとき、交差検証成績を少なくとも独立変数のひとつとして持つ関数を最適化させるように説明変数を逐次追加・除外することによって、説明変数を選抜して、良好な相関モデルを得る。このようなアプローチの優位性は、下記の実施例から推測されるように、次のとおりである。

１）病気や生体現象の背後で働いている重要な遺伝子やメカニズムを推定／特定でき、理解が深まる。

２）重要な遺伝子産物や細胞内物質だけに絞った廉価な診断用材料（ＤＮＡチップ、抗体チップなど）の設計が可能になる。

【００３７】

本実施形態では、交差検証成績を少なくとも独立変数のひとつとして持つ関数を最適化するように説明変数を段階的に取捨選択するが、たとえば具体的には、ステップワイズ(step wise)法に代表される説明変数を選択する選択手段と、リーブ・ワン・アウト(leave-one-out)法に代表される交差検証法に部分最小自乗法を適用して計算する計算手段と、前記計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを組合せて用いる。すなわち、ｍ個の説明変数の中から１以上の説明変数を選択し、次いで、部分最小自乗法を実行して交差検証成績を計算し、さらに、該計算結果を評価して、選択した説明変数の採用、不採用を判定する。この評価判定では、計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、少なくとも当該誤差の代表値を独立変数として持つ関数である当該誤差の誤差の代表値の単調減少関数の値が小さくなった場合に説明変数の取捨選択を判定する。このように、選択手段と計算手段と評価判定手段とを用いて、少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数を改善し続けて、その改善がみられなくなるまで改良し、部分最小自乗法モデルを決定する。なお、本実施形態では、サンプルを１個づつ逐次除外している(リーブ・ワン・アウト法)が、その代わりに、複数のサンプルを除外して交差検証成績を評価してもよ

い(リーブ・ｎ・アウト法)し、また、Khan et al.により用いられた３分割法(three-fold)等の他の方法を用いることもできる。３分割法では、説明変数をランダムにシャッフルして３つのグループに分ける。その中の２つのグループを用いてモデルを構成し、残りの１つのグループでモデルを評価する。また、説明変数の選択方法としてはステップワイズ法、非線形アルゴリズム(たとえば遺伝的アルゴリズムなど)を用いてもよく、変数選択に関して予め何らかの条件が分っていれば、それに応じて探索範囲を限定できる。

【００３８】

次に、データの収集と解析について具体的に説明する。図１は、遺伝子発現解析システムを示す。データ収集のため、予めいくつかのサンプルについて診断指標（たとえば病気のタイプないし進行度合いを含む）を判定し、また、そのサンプルされたものから細胞液を獲得し、ＤＮＡチップを用いてその細胞液中の多くの遺伝子産物の発現の量を測定する。測定には、共焦点型レーザスキャナ（たとえばAffymetrix社、４２８アレイスキャナ）１０を用いる。吸光度によりｍＲＮＡの量が測定される。このデータ収集は公知の方法である。測定データは、コンピュータ１２に送られ解析される。コンピュータ１２は、ＣＰＵ１４を備えた通常の構成のコンピュータであり、それに接続される記憶装置（たとえばハードディスク装置）１６の記録媒体（たとえばハードディスク）には、測定データ１８や解析ソフト２０が格納される。この解析ソフト２０を用いてデータ１８が解析され、生体の状態と遺伝子発現の量などとの相関モデルが決定される。

【００３９】

なお、説明変数の選択と、交差検証法に部分最小自乗法を適用する計算とを複数のコンピュータで実行させてもよい。交差検証予測の計算を複数個のコンピュータに分散させることで計算を加速することができる。

【００４０】

図２は、コンピュータ１２により実行される、生体の状態と遺伝子発現の量などとの相関モデルを得るためのデータ解析ソフト２０のフローチャートを示す。ここでは簡単に説明するため、少なくとも部分最小自乗法モデルの交差検証成績を独立変数として持つ関数として-PRESSを採用しているが、発明の範囲を限定す

るものでなく、実施例２～５においては別の関数を採用している。まず、相関モデル作成用のデータを入力する（Ｓ１０）。データはたとえばＤＮＡチップを用いて収集したものである。入力データ（サンプル集合）は、それぞれ目的変数（たとえば診断指標）とｍ個（たとえば２０００個）の説明変数（たとえば遺伝子発現の量）からなる。また、場合によっては、上述のデータ(訓練集合)以外に、テスト集合のデータを入力する。ここでテスト集合とは交差検証の評価のためのデータ群を意味するのではなく、モデル決定が終了した後にモデルの予測力をテストするためのデータ群である。

【００４１】

まず、初期設定として、選択された説明変数の数を０とし、交差検証成績CVの最良値CV_0を$-\infty$とする（Ｓ１２）。次に、説明変数の選択を行う。まず、説明変数を指す番号ｉを１とし（Ｓ１４）、第ｉ変数（遺伝子発現の量）を仮に採用して（Ｓ１６）、部分最小自乗法を実行し、交差検証成績CVを計算する（Ｓ１８、図３参照）。ここで、リーブ・ワン・アウト処理を用いる。これは、たとえば５０個のサンプルからなる訓練集合において、１番から５０番の全てを順次１個づつ除いて残りの４９個のサンプルで予測した結果と、その時除いた１個の結果とを比較し、その誤差が大きい場合に、仮に選択した説明変数(第ｉ変数)が適していないと判断する手法である。もし、得られた成績CVが現在の最良値CV_0より最適化されれば（Ｓ２０でＹＥＳ）、第ｉ変数を採用し、かつ、成績CVを新らしい最良値CV_0に更新する（Ｓ２２）。しかし、得られた成績CVが最良値CV_0より大きくなければ（Ｓ２０でＮＯ）、第ｉ変数を採用しない（Ｓ２４）。そして、ステップＳ１４に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる（Ｓ２６でＮＯ）まで繰り返す。ここで、相関モデルに採用する説明変数については１つづつ段階的に増加(追加)または減少（除外）して成績CVを評価判定している。すなわち、全体としての合致度合いがよくなるように各説明変数を解析に加えるかどうかを逐次判定しながら、説明変数の取捨選択を行い、これを、全体としての合致度合いがよくならなくなるまで繰り返す。以上の処理で改善があると、ふたたびステップＳ１４の初め(i=１)に戻り、それまでに選択されている説明変数を基に、さらに説明変数の選択を繰り返す。なお、ここで

はモデルの予測力を判断するために、訓練集合とテスト集合とに予め分割しておいたデータ集合を用いてデータ解析しており、上述の解析は、訓練集合を用いて行なった結果であるので、この結果からテスト集合について予測を行い、実測データとの合致度を評価（Ｓ２８）している。このような評価は必ずしも必要でないが、予測力を判断するには有効である。

【００４２】

図３は、リーブ・ワン・アウト処理を含む交差検証成績CVの計算（図２、Ｓ１８）のフローチャートを示す。ここで、選択された変数について交差検証成績が計算される。まず、PRESSの初期値を０とする（Ｓ１８０）。次に、ｎ個の集合内のサンプルを指す番号ｊを１とし（Ｓ１８２）、第ｊサンプル以外のｎ－１個のサンプルで部分最小自乗法を実行し（Ｓ１８４）、第ｊサンプルの目的変数を予測する（Ｓ１８６）。差の自乗を計算してPRESSに加算する（Ｓ１９０）。次に番号ｊを１増加し（Ｓ１８２）、同様の処理をおこなう。これを番号ｊ＝ｎまで各サンプルについて繰り返す。得られたPRESSは、１個のサンプルを順次除外して計算した予測値と実測値との差の平方和であり、予測誤差を表わす量である。この予測残差自乗和PRESSの符号を変えたものを交差検証成績CVとする（Ｓ１９２）。

【００４３】

本実施形態では、交差検証法を用いて、入力変数（説明変数）を段階的に１つづつ追加・除外しながら、交差検証成績（CV=-PRESS）を最適化する。ここで、説明変数の段階的な追加・除外の内容を理解しやすくするため、以下で、さらに具体的に５つのモデル構築手法について説明する。これらは、説明変数の逐次的な選択の手順が異なる。

【００４４】

図４は、第１のモデル構築手法を示す。データ集合においてどの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後(ｍ番目)の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ１１８）を繰り返しながら

判定(Ｓ１２０)し、改善する場合にはその説明変数を追加する（Ｓ１１４～Ｓ１２４）。そのような改善と追加がなくなる（Ｓ１２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。

【００４５】

さらに詳しく説明すると、まず、初期設定として、選択された説明変数の数ＮＰを０とし、交差検証成績CVの最良値CV_0を－∞とする（Ｓ１１２）。次に、説明変数の選択を行う。まず、変数ｉを１とし（Ｓ１１４）、第ｉ変数を仮に採用する（Ｓ１１６）。ただし、第ｉ変数がすでに採用されていれば（Ｓ１１５でＹＥＳ）、ステップＳ１１４に戻る。次に、部分最小自乗法を実行し、交差検証成績CVを計算する（Ｓ１１８）。ここで、リーブ・ワン・アウト処理を用いる。もし、得られた成績CVが現在の最良値CV_0より最適化されれば（Ｓ１２０でＹＥＳ）、第ｉ変数を採用し、かつ、成績CVを新らしい最良値CV_0に更新する（Ｓ１２２）。しかし、得られた成績CVが最良値CV_0より大きくなければ（Ｓ１２０でＮＯ）、第ｉ変数を採用しない（Ｓ１２４）。そして、ステップＳ１１４に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる（Ｓ１２６でＮＯ）まで繰り返す。以上の処理で改善があると、ふたたびステップＳ１１４に戻り、新しいループを開始する。ここで、それまでに選択されている変数を基に、さらに変数の選択を繰り返す。こうして、データ集合を用いて選択された変数を用いた相関モデルが得られる。

【００４６】

図５は、第２のモデル構築手法を示す。この手法では、全ての説明変数が選択されている状態を初期状態とする（Ｓ２１２）。次に、１番目の説明変数から最後(ｍ番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ２１８）を繰り返しながら判定(Ｓ２２０)し、改善する場合にはその説明変数を除外する（Ｓ２１４～Ｓ２２４）。そのような改善と除外がなくなる（Ｓ２２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。

【００４７】

さらに詳しく説明すると、まず、初期設定として、選択された説明変数の数ＮＰをｍとし、交差検証成績CVの最良値CV_0を－∞とする（Ｓ２１２）。すなわち、すべての説明変数を選択する。次に、説明変数の選択を行う。まず、変数ｉを１とし（Ｓ２１４）、第ｉ変数を仮に除外する（Ｓ２１６）。ただし、第ｉ変数がすでに除外されていれば（Ｓ２１５でＹＥＳ）、ステップＳ２１４に戻る。部分最小自乗法を実行し、交差検証成績CVを計算する（Ｓ２１８）。ここで、リーブ・ワン・アウト処理を用いる。もし、得られた成績CVが現在の最良値CV_0より最適化されれば（Ｓ２２０でＹＥＳ）、第ｉ変数を除外し、かつ、成績CVを新らしい最良値CV_0に更新する（Ｓ２２２）。しかし、得られた成績CVが最良値CV_0より大きくなければ（Ｓ２２０でＮＯ）、第ｉ変数を除外しない（Ｓ２２４）。そして、ステップＳ２１４に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる（Ｓ２２６でＮＯ）まで繰り返す。以上の処理で改善があると、ふたたびステップＳ２１４に戻り、新しいループを開始する。ここで、それまでに選択されている変数を基に、さらに変数の選択を繰り返す。こうして、データ集合を用いて選択された変数を用いた相関モデルが得られる。

【００４８】

図６は、第３のモデル構成手法を示す。この手法は、第１と第２の手法の直列的な組合せである。まず、どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後（ｍ番目）の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を追加選択し、そのような改善と追加がなくなるまで１番目の説明変数から上記逐次判定操作を繰り返す（Ｓ１１４～Ｓ１２６）。次に、１番目の説明変数から最後（ｍ番目）の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を除外し、そのような改善と除外がなくなるまで、１番目の説明変数から上記逐次判定操作を繰り返す（Ｓ２１４～Ｓ２２６）。

【００４９】

図７は、第４のモデル構築手法を示す。この手法は、第３の手法の変形である。まず、どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後(m番目)の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ１１８）を繰り返しながら判定(Ｓ１２０)し、改善する場合にはその説明変数を追加選択する（Ｓ１１４～Ｓ１２４）。そのような改善と追加がなくなる（Ｓ１２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。次に、１番目の説明変数から最後(m番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ２１８）を繰り返しながら判定(Ｓ２２０)し、改善する場合にはその説明変数を除外する（Ｓ２１４～２２４）。そのような改善と除外がなくなる（Ｓ２２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。上記逐次判定追加改善ステップまたは上記逐次判定除外改善ステップで少なくとも一度改善があれば（Ｓ２２７でＹＥＳ）、ステップＳ１１２に戻り、上記操作(Ｓ１１２～Ｓ２２７)を繰り返す。これを改善がなくなる（Ｓ２２７でＮＯ）までおこなう。

【００５０】

図８は、第５のモデル構築手法を示す。この手法は、第１と第２のスキームの並列的な組合せである。どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後（m番目）の説明変数までの説明変数ごとに逐次、その説明変数が選択されていない場合にはその説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ１１８）を繰り返しながら判定（Ｓ１２０）し、改善する場合にはその説明変数を追加する（Ｓ１１４～Ｓ１２４）。また、選択する説明変数ごとに、その説明変数がすでに選択されている場合には、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ２１８）を繰り返しながら

判定（Ｓ２２０）し、改善する場合にはその説明変数を除外する（Ｓ２１６～Ｓ２２４）。そのような改善と追加または除外がなくなる（Ｓ１２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。

【００５１】

次に、第４のモデル構築手法(図７)を適用した場合を、表１のデータ集合を例として説明する。このデータ集合に対して、部分最小自乗法による解析を用いて相関モデルを求める。表１のデータでは、サンプルの数ｎは１０であり、また、説明を容易にするため、説明変数の数ｍは１９と少なくしている。表１において、p1は目的変数を表わし、p2～p20は説明変数を表わす。(ただし表１では、表示の便宜のため、p16以降のデータを省略している。)第４手法(図７)のステップＳ１１４、Ｓ２１４とは異なり、説明変数を表わすｉはp20からp2まで逆に逐次処理することとした。CV評価値としてここでは予測残差自乗和(PRESS)を採用した。PRESSが小さいほど、CV評価値はよい。初期状態では、採用された説明変数の数ＮＰは０であり、PRESS=∞ (CV_0=-∞)である。

【００５２】

【表１】

表１　１０個のサンプルのデータ

#	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15
1	0.713	0.105	0.782	0.425	0.164	0.023	0.696	0.543	0.333	0.691	0.336	0.668	0.017	0.061	0.5
2	0.133	0.009	0.071	0.002	0.793	0.872	0.092	0.391	0.63	0.241	0.517	0.369	0.166	0.841	0.1
3	0.545	0.193	0.765	0.334	0.109	0.538	0.578	0.652	0.38	0.501	0.729	0.91	0.865	0.389	0.8
4	0.752	0.915	0.472	0.999	0.798	0.363	0.622	0.487	0.353	0.967	0.778	0.484	0.517	0.982	0.0
5	0.9	0.407	0.534	0.816	0.806	0.42	0.572	0.957	0.12	0.696	0.833	0.051	0.377	0.849	0.4
6	0.455	0.587	0.721	0.53	0.252	0.434	0.882	0.486	0.741	0,243	0.893	0.947	0.462	0.952	0.2
7	0.427	0.652	0.515	0.426	0.764	0.592	0.595	0.595	0.551	0.606	0.416	0.163	0.316	0.718	0.6
8	0.042	0.902	0.274	0,899	0.402	0.469	0.668	0.945	0.746	0.912	0.97	0.515	0.368	0.514	0.4
9	0.935	0.276	0.936	0.101	0.54	0.356	0.899	0.71	0.924	0.792	0.486	0.329	0.501	0.076	0.5
10	0.54	0.021	0.505	0.224	0.724	0.431	0.071	0.968	0.482	0.322	0.773	0.543	0.353	0.107	0.9

【００５３】

【表２】

表２　　表１のデータについての１０の段階での変数選択結果

0			∞	-
1	追加	p20	0.111	p20
2	追加	p18	0.090	p18 & p20
3	追加	p16	0.073	p16 & p18 & p20
4	追加	p10	0.073	p10 & p16 & p18 & p20
5	追加	p6	0.062	p6 & p10 & p16 & p18 & p20
6	追加	p3	0.060	p3 & p6 & p10 & p16 & p18 & p20
7	追加	p12	0.055	p3 & p6 & p10 & p12 & p16 & p18 & p20
8	除外	p20	0.053	p3 & p6 & p10 & p12 & p16 &
9	除外	p10	0.050	p3 & p6 & p12 & p16 & p18
10	追加	p13	0.048	p3 & p6 & p12 & p13 & p16 & p15

【００５４】

先に述べたように、変数はp20からp2まで逆の順で処理する。表２は、表１のサンプルについて、左端の数字は、変数の取捨選択で改善がみられた１０の段階を示す。なお、０は初期状態を意味する。次の列の「追加」と「除外」は、追加のループと除外のループの処理であることを意味する。次の列の変数は、追加または除外された変数を示す。次の列は、交差検証成績(PRESSをサンプル数で割ったもの)を示す。右端の列は、その段階で選択されている変数を示す。

【００５５】

初期状態では、変数は全くない状態であり、PRESSは∞である。表２に示すように、最初、p20を説明変数として採用すると、PRESS=0.111となり、初期値に比べて改善されるので、説明変数p20の追加を実施する。次に、変数p19を加えてp19とp20の２つを説明変数とすると、PRESS=0.129となり改善をもたらさないので、p19は追加しない。次に、説明変数p18を加えるとPRESS=0.090となり、改善するので、p18を追加し、p18とp20を説明変数とする。以下同様に表２に示すように続く。(ここで、p10を追加採用するのは、小数点以下４桁目で改善されているた

めである。)説明変数p20～p2の１回目のループを終了した時点で、説明変数がp3、p6、p10、p16、p18およびp20となり、PRESS=0.60となる。２回目のループでは、説明変数p12が追加され、PRESS=0.55となる。３回目のループでは追加による改善がなく、ひとまずＳ１１４～Ｓ１２６の追加処理を終了し、Ｓ２１４に移る。この時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況は表３のとおりである。

【００５６】

表３は、１０のサンプルについて、表２の７で示す段階まで処理が進んだ時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況を示す。ここで、モデル予測とリーブ・ワン・アウト予測のそれぞれにおいて、計算値と実測値との誤差を示す。さらに、その下側に、誤差の自乗平均、相関係数Rの自乗および予測相関係数Qの自乗を示す。

【００５７】

【表３】

表３　表２の段階７での処理結果

		モデル予測値		リーブワンアウト予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.757	-0.044	0.693	0.020
2	0.133	-0.056	0.189	-0.051	0.184
3	0.545	0.497	0.048	0.480	0.065
4	0.752	0.646	0.106	0.495	0.257
5	0.900	0.687	0.214	0.557	0.343
6	0.455	0.489	-0.034	0.512	-0.057
7	0.427	0.624	-0.198	0.672	-0.245
8	0.042	0.349	-0.307	0.517	-0.475
9	0.935	0.865	0.070	0.782	0.153
10	0.154	0.197	-0.044	0.285	-0.132
	0.093		0.024		0.055

R^2=0.744　　Q^2=0.407

【００５８】

次に、Ｓ２１４から始まる除外処理の１回目のループにおいて、説明変数p10とp20を除外することが改善をもたらした。２回目のループでは改善がなく、Ｓ２１４～Ｓ２２６を終了するが、Ｓ２２７の判断により再度Ｓ１１２に戻る。次に、追加処理の１回目のループにおいて、p13の追加だけが改善をもたらしたが、続く除外処理の１回目のループでは、改善がなかった。もう一度Ｓ１１２に戻り、ステップＳ１１４～Ｓ１２６およびステップＳ２１４～Ｓ２２６では改善がなくなったのを確認して、処理を終了した。こうして選択された説明変数は、p3、p6、p12、p13、p16およびp18の５個であり、PRESS=0.048となった。詳細は表４のとおりである。

【００５９】

表４は、表２の段階１０まで処理が進んだ時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況を示す。

【００６０】

【表４】

表４　表２の段階１０での処理結果

		モデル予測		リーブ・ワンアウト予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.771	-0.058	0.663	0.050
2	0.133	-0.013	0.146	0.041	0.092
3	0.545	0.610	-0.065	0.595	-0.050
4	0.752	0.524	0.228	0.380	0.372
5	0.900	0.696	0.205	0.543	0.357
6	0.455	0.591	-0.137	0.623	-0.168
7	0.427	0.638	-0.211	0.696	-0.269
8	0.042	0.189	-0.147	0.268	-0.226
9	0.935	0.841	0.094	0.756	0.179
10	0.154	0.209	-0.055	0.294	-0.140
	0.093		0.022		0.048
		R^2=0.765		Q^2=0.482	

【００６１】

なお、説明変数の数が多い時に強いとされる部分最小自乗法であるが、p20～p2の全てを説明変数として採用した場合には、表５に示すように、PRESS=0.124となった。すなわち、リーブ・ワン・アウト処理は、平均値からの誤差(0.093)よりも悪い成績をもたらす。

【００６２】

【表５】

表５　全ての説明変数を採用した場合の処理結果

		モデル予測		リーブワンアウト予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.712	0.001	0.527	0.186
2	0.133	-0.073	0.206	0.222	-0.090
3	0.545	0.561	-0.016	0.538	0.007
4	0.752	0.656	0.096	0.351	0.402
5	0.900	0.691	0.209	0.432	0.469
6	0.455	0.519	-0.064	0.562	-0.107
7	0.427	0.583	-0.156	0.629	-0.203
8	0.042	0.430	-0.388	0.724	-0.682
9	0.935	0.794	0.140	0.480	0.454
10	0.154	0.182	-0.029	0.457	-0.303
	0.093		0.029		0.124
		R^2=0.684		Q^2=-0.330	

【００６３】

【実施例】

次に、実施例を挙げて本発明をさらに詳細に説明するが、本発明はこれらの例によって何ら限定されるものではない。

【００６４】

実施例１：　部分最小自乗法の交差検証成績を考慮した特徴抽出によるＤＬＢＣＬ患者のデータ解析

Ｐ．Ｏ．Brownらのホームページ（http://llmpp.nih.gov/lymphoma/）より入手した２８名のＤＬＢＣＬ（リンパ腫）患者のデータを、２０名のデータからなる訓練集合と８名のデータからなるテスト集合に分けた。目的変数に生存月数を採用し、説明変数には１８４３２スポットのうち、２８データにおいてch1、ch2と

もに正の数となる１２８３２スポットのlog(ch1/ch2)値を採用した。

【００６５】

訓練集合において部分最小自乗法（PLS）のモデル決定を試みた。１２８３２変数全てを用いて部分最小自乗法の解析をしたところ、リーブ・ワン・アウト予測は有意($Q^2 > 0.5$)にはならなかった。次にリーブ・ワン・アウト予測誤差が最小になるように説明変数を段階的に１つづつ増減した。モデル構成手法としては前述の第３のモデル構成手法において説明変数の追加及び除外の順番並びにリーブ・ワン・アウト処理におけるサンプルの除外の順番が異なるほかは同様な方法を用いた。すなわち、どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、最後（ｍ番目）の説明変数から最初（１番目）の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理（ここでは、最後（ｎ番目）のサンプルから最初（１番目）のサンプルを逐次除外した）を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を追加選択し、そのような改善と追加がなくなるまでｍ番目の説明変数から上記逐次判定操作を繰り返す（Ｓ１１４～Ｓ１２６）。次に、最後（ｍ番目）の説明変数から最初（１番目）の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理〔ここでも最後（ｎ番目）のサンプルから逐次除外した〕を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を除外し、そのような改善と除外がなくなるまで、最後（ｍ番目）の説明変数から上記逐次判定操作を繰り返す（Ｓ２１４～Ｓ２２６）。その結果、有意なモデル（$R^2=0.988$、$Q^2=0.895$、ＮＰ＝342）を得た。図９は、このデータについての最小自乗法成績を示す。図９において、ひし形は訓練集合のデータ（20人）を示し、三角は、それらについての交差検証成績のデータを示す。また、四角はテスト集合のデータ（８人）を示す。得られた部分最小自乗法モデルは、テスト集合のうち、４/８をきわめて良好に、また１/８を良好に予測するものであった。

【００６６】

なお、上述の多変量解析によるデータ解析では、扱ったサンプルはＤＮＡチップを用いて得たデータであった。しかし、このデータ解析は、ＤＮＡチップを用いて得たデータに限定されるものではなく、蛋白質発現量、細胞内物質の量などのデータに対しても有用であろうことは容易に推測されることである。

【００６７】

以下の実施例２～７では、部分最小自乗法を用いて選抜した少ない個数の説明変数について、通常の統計的手法または多変量解析手法（比例ハザード法、重回帰分析、適応最小自乗法、ロジスティック回帰分析法、線型判別分析法など）を適用する。

【００６８】

実施例２：　部分最小自乗法の交差検証成績を考慮した特徴抽出と比例ハザード解析による２４０名のＤＬＢＣＬ患者の生存時間解析

RosenwaldらがWeb上（http://llmpp.nih.gov/DLBCL/）で公開している２４０名のＤＬＢＣＬ(びまん性大細胞型Ｂリンパ腫)のデータセットをダウンロードして用いた。全データを訓練集合として利用した。スポットパターンでχ１またはχ２が０となるものを除いた７３９９スポットについてlog(χ１/χ２)を計算して説明変数とした。本実施例では実施例１と異なり、生存時間として観測打切り時間と死亡時間とが混在していることを考慮してカプラン・マイヤー（Kaplan-Meier）法による生命表を適用して事象発生時点での生存確率(P_{KM})を求め、ロジット変換($\log(P_{KM}/1-P_{KM})$)した値を目的変数とした。カプラン・マイヤー法による生命表は集団としての生存確率を示すが、ここでは、個人ｊを含む集団としての事象発生時点での残存確率（変化の発生しなかったものが残存する確率）を個人ｊの事象発生時点での残存時間に読み代えるという新規な考え方を用いている。また、この確率をロジット変換して、変化の発生傾向を表現するロジット値に変換して、目的変数とした。訓練集合内の交差検証はリーブ・ワン・アウト法によって行ない、$PRESS\times1.02^{NP}$が小さくなるようにパラメータを逐次取捨選択して部分最小自乗法モデルを得た。ここで、交差検証成績（CV=-PRESS）の代わりに、少なくとも交差検証成績を独立変数として持つ関数の１つである関数$-PRESS\times1.02^{NP}$を改善して部分最小自乗法モデルを得た。ここでPRESSはリーブ・

ワン・アウト予測の残差自乗和であり、ＮＰは、選択された説明変数の数である。

【００６９】

図７のフロー中の交差検証成績CVを-PRESS×1.02^{NP}と読み換えて、処理を実行することにより、下記の１９個の遺伝子の発現が説明変数として選抜された。ここでdata IDはWebデータ元でのID番号を示す。またACCESSIONはGenBankのアクセション番号であり、アクセション番号の無い行はデータ元でのみ明らかとなっている遺伝子（Unknown）ないしESTであり、論文記載の方法によって入手することができる。

【００７０】

ACCESSION	data ID	comment
U03398	#(27876)	tumor necrosis factor (ligand) superfamily, member 9
M65066	#(27394)	protein kinase, cAMP-dependent, regulatory, type I, beta
--	#(27104)	(Unknown)
AK001546	#(25048)	Homo sapiens cDNA FLJ10684 fis, clone NT2RP3000220
--	#(31372)	(Unknown)
U15085	#(28178)	major histocompatibility complex, class II, DM beta
BC003536	#(24983)	hypothetical protein MGC10796
--	#(16113)	(Unknown)
M23452	#(16822)	small inducible cytokine A3
	#(24433)	(Unknown)
X00437	#(27480)	T cell receptor beta locus
U12979	#(24377)	activated RNA polymerase II transcription cofactor 4
X52479	#(17773)	protein kinase C, alpha

H96306	#(16578)	bone marrow stromal cell antigen 1
U70426	#(19255)	regulator of G-protein signaling 16
AA830781	#(33358)	EST
AA804793	#(25022)	EST
H57330	#(26383)	EST
S69790	#(27184)	WAS protein family, member 3

【００７１】

これらの遺伝子の発現を説明変数の候補として比例ハザード(hazard)解析を試みた。比例ハザード法とは、生存率の解析に時間を考慮した統計的手法である。解析の実行はプログラムパッケージＪＭＰ(JMP Sales SAS Campus DriveCary, NC 27513 USA)を用いて行なった。変数削除基準としてP≧0.05を採用した変数減少法によって更に絞り込んだ結果、１４遺伝子の発現からなる以下の比例ハザード式が得られた。ここでGenbank（ジーンバンク）のアクセション番号ないしdata IDで示される各項は、各遺伝子のlog(χ１/χ２)値であり、またＰは統計的な有意性が成り立たない危険率である。この式の右辺から求められるハザード(hazard)値が大きいほど、死亡傾向が大きい。

【００７２】

hazard = 0.370 #(27104) +0.589 AK001546 -0.366 #(31372) -0.276 U15085
-0.307 #(16113) +0.409 M23452 -0.350 #(24433) -0.297 X00437
+0.321 U12979 -0.585 X52479 -0.457 U70426 +0.561 AA830781
-0.430 H57330 +0.433 S69790

P<0.0001

【００７３】

Rosenwaldらは、単相関の比例ハザード解析を行なって、５群(１７遺伝子)の診断指標を選抜している。図１０に、本実施例で得られたハザード値（Hazard)とRosenwaldらの診断指標がどの程度、生存時間を説明できているかを比較した。Rosenwaldらの５群のパラメータを同時に用いた比例ハザード式ではProliferationパラメータがP>0.05で統計的に有意でないなどの問題を有していため、これを除く４群のパラメータを同時に含めたハザード値も比較のために掲載した。こ

こで、菱形は死亡した人または打ち切った人のデータを示し、四角は生存している人のデータを示す。

【００７４】

これらの診断指標のうち、本実施例で求めたハザード値と生存時間との相関は際立って明白である。即ちハザード値は生存時間につれて減衰しており、大きなハザード値の患者は長く生きることが出来ないことが示されている。一方、Rosenwaldらの指標はいずれも生存時間を診断するには不十分なものである。数百、数千という数のパラメータの中から効率的に最適のパラメータセットを見出すことは比例ハザード解析だけではできないことである。しかし以上のようにカプラン・マイヤー法、ロジット変換、部分最小自乗法の交差検証成績を考慮した特徴抽出、比例ハザード解析を組み合わせることで、従来に無い、有効な診断指標を得ることができた。統計学的に異質なモデルをこのように組み合わせることによってこのような良好な結果が得られたことは意外でもあり、興味深いことであった。患者の生存時間を予測することは、QOLを含めた治療計画や人生設計などを判断する上で重要な情報をもたらすものであり、本実施例で求められた診断モデルは社会的に価値のあるものである。

【００７５】

また、変数削除基準としてP≧0.001を採用した変数減少法によって更に絞り込むと、７遺伝子の発現からなる以下の比例ハザード式が得られた。このように、変数削除基準を変えることにより、選択される説明変数の数を制御できる。

【００７６】

hazard = -0.426 U15085 +0.350 M23452 -0.521 X52479
-0.450 U70426 -0.586 H57330 +0.476 S69790

【００７７】

図１１は、右辺を計算して求められるハザード値を縦軸とし、生存時間を横軸としたプロットを示す。図１０と同様に、図１１において、菱形は死亡した人または打ち切った人のデータを示し、四角は生存している人のデータを示す。

【００７８】

実施例３：　部分最小自乗法の交差検証成績を考慮した特徴抽出と比例ハザー

ド解析による４０名の乳癌患者の生存時間解析

SorleらがWeb上(http://genome-www.stanford.edu/breast_cacer/mopo_clinical/)で公開している乳癌患者のデータセットをダウンロードして用いた。全データを訓練集合として利用した。データセットの大部分は、タイプＡ，Ｂという２種類のＤＮＡチップで測定されたそれぞれ４０名、２４名の患者よりなるが、ここではタイプＡのデータを用いた。生存時間データより実施例２と同様にロジット値を求め、目的変数とした。説明変数としては、データに欠測のある遺伝子を除いた６８９１件のLOG_RAT2N_MEAN値を採用した。そして、少なくとも交差検証成績を独立変数として持つ関数の１つである、交差検証成績と説明変数ＮＰの関数PRESS×1.13^{NP}が小さくなるようにパラメータを逐次取捨選択して部分最小自乗法モデルを得た。図７のフロー中の交差検証成績CVを-PRESS×1.13^{NP}と読み換えて、処理を実行することにより、下記の１０個の遺伝子の発現が説明変数として選抜された。

【００７９】

ACCESSION	comment
AA406242	(guanosine monophosphate reductase)
AA598572	(spleen tyrosine kinase)
H73335	(Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 980547)
W84753	(Homo sapiens cDNA FLJ13510 fis, clone PLACE1005146)
AA703058	(myeloperoxidase)
N71160	(cytochrome c oxidase subunit Vib)
AA453345	(a protein tyrosine kinase)
AA054669	(Homo sapiens, clone IMAGE:3611719, mRNA, partial cds)
N32820	(ESTs, Weakly similar to ALU1_HUMAN ALU SUBFAMILY J SEQUENCE CONTAMINATION WARNING ENTRY [H. sapiens])
R05667	(suppressor of potassium transport defect 3)

【００８０】

これらを説明変数の候補として、比例ハザード解析において変数削除基準としてP≧0.05を採用した変数減少法を試み、７遺伝子の発現からなる以下の比例ハザード式が得られた。ここでアクセッション番号で示される各項はそれぞれの遺伝子のLOG_RAT2N_MEANである。

【００８１】

hazard = -0.821 AA406242 +1.556 AA598572 -1.074 H7335 +1.418 W84753
-1.290 AA703058 +2.182 N71160 +0.828 AA453345

P<0.0001　変数のP<0.05

【００８２】

図１２に、右辺を計算して求められるハザード値を縦軸とし、生存時間を横軸としたプロットを示す。ここでもハザード値が優れた診断指標となることが示されている。図１２において、菱形は死亡した人または打ち切った人のデータを示し、四角は生存している人のデータを示す。

【００８３】

変数削除基準としてP≧0.001を採用した変数減少法によって更に絞り込んだ。これにより、３遺伝子の発現からなる以下の比例ハザード式が得られた。このように、変数削除基準を変えることにより、説明変数の数を制御できた。

【００８４】

hazard = 1.453 AA598572 -1.473 AA703058 +1.071 AA453345

【００８５】

図１３は、右辺を計算して求められるハザード値を縦軸とし、生存時間を横軸としたプロットを示す。ここで、菱形は死亡した人のデータを示し、四角は生存している人のデータを示す。

【００８６】

実施例４：　部分最小自乗法の交差検証成績を考慮した特徴抽出と重回帰分析による４０名の乳癌患者の再発予測解析

SorleらのDNAチップＡで６８９１遺伝子の発現が測定された４０名の患者をデータセットとして用いた。再発の有無を目的変数として、PRESS×1.10^{NP}が小

さくなるようにパラメータを逐次取捨選択して１１遺伝子の発現からなる部分最小自乗法モデルを得た。

【００８７】

ACCESSION	comment
AA434397	integrin, beta 5
T83209	ESTs
N53427	KIAA1628 protein
N29639	cytidine monophosphate-N-acetylneuraminic acid hydroxylase
AA485739	major histocompatibility complex, class II, DR beta 5
AA425861	enoyl Coenzyme A hydratase 1, peroxisomal
H84871	Ste-20 related kinase
T64312	prostate cancer overexpressed gene 1
T59518	solute carrier family 2, (facilitated glucose transporter) member 8
AA406231	KIAA0381 protein
AA037488	prolactin

【００８８】

次に、選抜された遺伝子発現を説明変数とし、再発の有無を目的変数として、通常の多変数解析法の一つである重回帰分析によって判別分析を実行した。解析の実行はプログラムパッケージＪＭＰを用いて行なった。変数削除基準としてP≧0.15を採用した変数減少法によってさらに絞り込んだ結果、１０遺伝子の発現からなる以下の重回帰式が得られた。この式で計算されるOLS値が正の時は再発の可能性が高く、負の時は低い。

【００８９】

OLS = -0.215 AA434397 +0.227 T83209 -0.209 N53427 +0.139 N29639
+0.165 AA485739 +0.133 AA425861 -0.084 H84871 -0.193 T64312
+0.237 T59518 +0.176 AA037488 -0.278

R^2=0.84797、　判別正解率 97.5%

【００９０】

上式に含まれる各パラメータをそれぞれ１つ用いて判別分析式を作成した場合のＰ値及び決定係数を以下の表６に示す。

【００９１】

【表６】

表６

Accession No.	P value	決定係数(R^2)
AA434397	0.0334	0.090273
T83209	0.0601	0.066005
N53427	0.0004	0.268678
N29639	0.0552	0.069483
AA485739	0.0421	0.080733
AA425861	0.0861	0.05122
H84871	0.087566	0.087566
T64312	0.0004	0.263207
T59518	0.0066	0.157196
AA037488	0.0031	0.187627

単独では有意とはならない(P>0.05)パラメータが３つ存在し、また、どのパラメータも決定係数が小さい。従って、パラメータを1つずつ吟味するだけでは、上式のような良好な判別式は得られなかった。また数百、数千という数のパラメータの中から効率的に最適のパラメータセットを見出すことは重回帰分析だけではできないことである。しかし、以上のように、部分最小自乗法の交差検証成績を考慮して特徴抽出することにより、従来に無い、有効な診断指標を得ることができた。乳癌の再発可能性を予測することは、QOLを考慮した治療計画を立案し判断するうえで、社会的に求められているところのものである。

【００９２】

実施例５：　部分最小自乗法の交差検証成績を考慮した特徴抽出と適応最小自

乗法による４０＋２４名の乳癌患者の再発予測解析

ＤＮＡチップのタイプＡ(４０名)とタイプＢ(２４名)に共通する３４４８遺伝子に限って解析を試みた。PRESS×1.17^{NP}が小さくなるようにパラメータを逐次取捨選択して部分最小自乗法モデルを得た。選抜された遺伝子発現を説明変数とし、適応最小自乗法によって判別分析を実行した結果、次式が得られた。次式で計算されるＡＬＳ値が０．５より大きいと再発の危険性が存在する。

【００９３】

ALS = 0.31 H11482 -0.29 T64312 -0.32 AA045340 +0.01

R^2 = 0.65, eps = 0.13, 判別正解率 90.0%

【００９４】

下記の表７にみるように、H11482 は単相関では有意ではなく、他の変数と同時に用いることで初めて把握できたパラメータである。また、表８は、上式を用いてタイプＢの患者を予測した結果である。本判別式の感度＝81.8%、特異度＝53.8%となり、χ^2＝3.233(5%<P<10%)、予測判別正解率＝66.7%、という統計的に有意な結果を得た。タイプＡ、ＢはＤＮＡチップの構成の相違に基づく測定誤差が存在すると思われるデータであるにもかかわらず、タイプＡで訓練したモデルでタイプＢの予測に危険率１０%以下で成功したことは勇気付けられる結果である。

【００９５】

また、PRESS×1.12^{NP}が小さくなるように選んだ場合には、以下の遺伝子の発現を説明変数とする部分最小自乗法モデルを得た。

【００９６】

H11482、T64312、R99749、T65211、AA427625、AA455506

【００９７】

これらを説明変数の候補として、リーブ・ワン・アウトを指標にして、さらに絞り込んだ結果、次の判別式を得た。

ALS = 0.53 H11482 -0.31 T64312 -0.33 R99749 -0.26 AA455506 +0.10

R^2 = 1.00, eps = 0.10, 判別正解率 100.0%

【００９８】

パラメータを1つずつ吟味するだけでは、上式のような良好な判別式は得られなかった。また数百、数千という数のパラメータの中から効率的に最適のパラメータセットを見出すことは、適応最小自乗法、ロジスティック回帰分析、その他の判別分析手法だけではできないことである。しかし、以上のように、部分最小自乗法の交差検証成績を考慮して特徴抽出することにより、従来に無い、有効な診断指標を得ることができた。

【００９９】

【表７】

表７　パラメータの交絡作用

パラメータ	R	Nmis(/40)
H22482	0.361	14
T64312	0.607	8
AA045340	0.572	9
T64312 & AA045340	0.716	6
H11482 & T64312 & AA045340	0.804	4

【０１００】

【表８】

表８　タイプＢの２４患者の予測

観察値	予測値	頻度
－	－	7
＋	－	2
－	＋	6
＋	＋	9

【０１０１】

実施例６

部分最小自乗法の交差検証成績を考慮した特徴抽出とロジスティック回帰分析法または線型判別分析法による４０＋２４名の乳癌患者の再発予測解析

実施例５での１つめの適応最小自乗法による解析をロジスティック回帰分析法に置き換えた場合、次の判別式が得られた。

【０１０２】

LORA = 7.92 H11482 -5.69 T64312 -6.41 AA045340 -9.73

R^2 = 0.63, x2 = 35.00 (P<0.0001), 判別正解率 90.0%

【０１０３】

右辺で求められるＬＯＲＡ値が正の場合には再発の危険性が存在する。係数の比率や相関係数は実施例５の適応最小自乗法の場合と異なるものの、各患者の識別結果は全く同一であった。またタイプＢの患者を予測した結果も表７と同じになった。

【０１０４】

次に、実施例５での適応最小自乗法による解析を線型判別分析に置き換えて解析して、次の判別式が得られた。

【０１０５】

LDA = 2.45 H11482 -2.35 T64312 -2.56 AA045340 -4.03

判別正解率 80.0%

【０１０６】

右辺で求められるＬＤＡ値が正の場合には再発の危険性が存在する。係数の比率や相関係数は、実施例５の適応最小自乗法の場合と異なり、各患者の識別結果も若干異なったが、概ね同一であった。また、タイプＢの患者を予測した結果も表７と同じになった。

【０１０７】

以上の実施例４，５，６では、乳癌の再発の有無を目的変数としている。したがって、部分最小自乗法の交差検証成績を考慮して特徴抽出する方法が、目的変数が名義尺度や順序尺度などのデータである場合にも有効であることが示された。なお、名義尺度とは、対象（サンプル）をある分類に属するかどうかを測り分けるときの分類で、分類の間に大小や順序はない。また、順序尺度とは、対象の特定の分類について測り分けるときの分類であり、分類の間に大小、高低といった順序がある。

【０１０８】

実施例７

部分最小自乗法の交差検証成績を考慮した特徴抽出と比例ハザード解析による４０名の乳癌患者の再発時間解析

実施例４と同じデータを用いて、再発の時系列データを基に実施例２と同様の方法で求めたロジット値を目的変数として、PRESS×1.15^{NP}が小さくなるようにパラメータを逐次取捨選択して９遺伝子の発現からなる部分最小自乗法モデルを得た。これらの遺伝子発現の測定値を説明変数として比例ハザード解析において変数削除基準としてP≧0.05を採用した変数減少法を試み、８遺伝子からなる、以下の比例ハザード式が得られた。

【０１０９】

hazard = 1.122 AA448641 -1.781 R78516 -1.434 R05934 +2.165 W84753
-1.923 AA629838 +2.665 H08581 +1.875 AA045730 +1.269 AI250654

P<0.0001

【０１１０】

図１４は、右辺を計算して求められるハザード値を縦軸とし、再発時間を横軸としたプロットを示す。ここで、菱形は再発しない人のデータを示し、四角は再発している人のデータを示す。ここでもハザード値が優れた診断指標となっており、生存時間に限らず、時間とともに確率的に発生する生体の状態の変化を解析する手法として、本発明の手法が有効であることが示されている。

【０１１１】

変数削除基準としてP≧0.005を採用した変数減少法によって更に絞り込んだ場合には、４遺伝子の発現からなる以下の比例ハザード式が得られた。

【０１１２】

hazard = 1.559 W84753 +2.265 H08581 +1.473 AA045730 +1.237 AI250654

【０１１３】

図１５は、右辺を計算して求められるハザード値を縦軸とし、再発時間を横軸としたプロットを示す。ここで、菱形は再発しない人のデータを示し、四角は再発している人のデータを示す。

【０１１４】

実施例８： Genbankアクセッション番号H11482、T64312、AA045340を含む乳癌再発性診断用ＤＮＡチップの作成と測定

実験医学別冊「ゲノム機能研究プロトコール」(ISBN4-89706-932-7 C3047) p34-38記載の関直彦、永杉友美、東孝典、吉川勉、鈴木収、村松正明らの方法に順じてＤＮＡチップの作成と測定を行なう。Genbankアクセッション番号H11482、T64312、AA045340のｃＤＮＡを用いる。

【０１１５】

プローブ用の各ＰＣＲ産物をエタノール(和光純薬，Cat#057-00456)で沈殿させ、2μg/μlとなるようにＤＤＷで調整する。ニトロセルロース(GibcoBRL Cat#41051-012)４mg/mlのＤＭＳＯ溶液を等量加え、よく混和させて１００℃で５分間熱変性を行ない、氷上で急冷する。次いで室温に戻し、ＤＮＡスポッターSPBI02000(日立ソフトエンジニアリング)を用いてカルボジイミドスライドガラス(日清紡)へのスポッティングを速やかに行なう。スポットの乾燥を確認し、Ultraviolet crosslinker(アマシャムファルマシアバイオテック社)を用いて６０mJ/cm^2で紫外クロスリンク処理を行ない、ガラスラックに立てて室温保存する。

3%BSA、0.2M NaCl、0.1M Tris(PH 7.5)、0.05% Triton X-100よりなるブロッキング液に上記マイクロアレイを浸け、約３０分間放置する。次いで、ガラスに付着している溶液をよく切り、３７℃で乾燥させる。TEバッファー(PH 8.0，ニッポンジーン Cat #316-90025)で３回軽く洗い、プレートホルダーに入れて軽く遠心(1000 rpm，１分間)して余分な水分を除去する。

【０１１６】

次に、乳腺正常株SV-40及び乳癌細胞株MCF-7、MDA-MB-468又はT-47-Dの各細胞液より、TRIZOL(GibcoBRL，Cat#15596-018)、Oligotex dT30<Super> (TaKaRa，Cat#W9021A)を用いてマニュアルに従って、mRNAを精製する。mRNA ２μgを6.4μlのＤＥＰＣ処理し、ＤＤＷに溶かし、Oligo dTプライマー 9μl、5X SuperScript IIバッファー(GibcoBRL，Cat#18089-011) 6μl、DTT(SuperScriptの付属) 3μl、50 X dNTP 0.6μl、Cy3-dUTP(アマシャムファルマシアバイオテク Cat# PA53022)又はCy5-dUTP (アマシャムファルマシアバイオテク Cat# PA55022) ３μl

、SuperScript II ２μlよりなる溶液を加え、４２℃で２時間反応させる。途中１時間経過時点で、SuperScript IIを１μlを追加する。1.5μlアルカリバッファー(1N NaOH / 20nM EDTA)を加え、65℃で１０分間反応させ、TEバッファーを270μl、1N HClを1.5μl加えて、Cy3，Cy5ラベルの反応液を２つ纏めて１本のMicrocon-YM-30(Millipore/Amicon，Cat#42410)に移す。１０，０００rpmで上のカップに残る液量が約１０μlになるまで遠心を続け、カップを通りぬける液を別のチューブに移し替え、その後、上のカップにTE バッファー５００μl、Human Cot-1 DNA(GibcoBRL Cat#15279-011) ２０μgを加え、再び液量が１０μl以下になるまで遠心を続ける。３，０００ rpmで３分間遠心し、蛍光標識したＤＮＡを回収する。DDWとyeast RNA(Sigma，Cat#R7125) 50μg、poly(A) (ロッシュダイアグノスティクス，Cat#108 626) 50μgを加えて20μlにし、PCR用のチューブに移し換え、さらに4.25μl 20 X SSC(GibcoBRL，Cat#15553-035)と0.75μl 10% SDS(GibcoBRL，Cat#15553-035)を加え、PCR用の機器で１００℃、１分間熱変性させ、次いで、室温で30分間放置して、ゆっくり冷却する。

【０１１７】

蛍光標識したＤＮＡの全量をカバーガラスにのせ、泡が入らないように注意しながら前記マイクロアレイにかぶせ、水で濡らしたキムタオルを底に敷いたハイブリダイゼーションチェンバーに入れて密閉する。毎分２～４サイクルで軽く振とうさせながら、６５℃で一晩ハイブリダイズする。ハイブリダイゼーションチェンバーからマイクロアレイを取り出し、カバーガラスがのったままの状態で静かに2 X SSC / 0.1% SDS溶液中に入れ、５分間シェイキングし、カバーガラスが自然にはがれるのを待つ。カバーガラスがはがれたところでマイクロアレイをスライドガラスラックに入れ、もう一度2 X SSC / 0.1% SDS溶液中で５分間軽く振とうして洗う。さらに0.2 X SSC / 0.1% SDS ４０℃で５分間２回洗い、0.2 X SSCでリンスする。マイクロアレイを別の乾いたプレパラートケースに移し、マイクロタイタープレート用の遠心機で軽く遠心して(1000 rpm，１分室温)マイクロアレイ上の水分を除く。ScanArray4000(GSI luminonics社)でシグナルを読み込み、解析ソフトにはQuant Array (GSI liminonics社)およびChip Space(日立ソフトウェアエンジニアリング)を用いる。

【０１１８】

【発明の効果】

生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するとき、説明変数の選択と交差検証法とを用いて変数を絞り込むことができる。これにより、良好でかつ予測力のある多変量解析モデル（相関モデル）が得られる。特に遺伝子発現の量のように、説明変数の数がたとえば１０００以上と膨大な場合に有用である。変数の数を少なくすることにより、病気や生体現象の背後で働いている重要な遺伝子やメカニズムを推定／特定でき、理解が深まる。また、重要な遺伝子産物や細胞内物質だけに絞った廉価な診断用材料（ＤＮＡチップ、ＤＮＡ含有ベクター、抗体チップなど）を設計し、提供できる。

【０１１９】

また、時間とともに確率的に発生する生体の状態の変化から導出された量を目的変数として用いて、時間とともに確率的に発生する生体の状態の変化と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定できる。

【０１２０】

また、部分最小自乗法を用いて説明変数の個数を少なくすると、通常の統計的手法または多変量解析手法が適用可能になる。

【図面の簡単な説明】

【図１】　遺伝子発現解析システムのブロック図

【図２】　解析ソフトのフローチャート

【図３】　交差検証成績CVの計算のフローチャート

【図４】　変数選択の第１モデル構築手法のフローチャート

【図５】　変数選択の第２モデル構築手法のフローチャート

【図６】　変数選択の第３モデル構築手法のフローチャート

【図７】　変数選択の第４モデル構築手法のフローチャート

【図８】　変数選択の第５モデル構築手法のフローチャート

【図９】　最小自乗法モデルの成績を示すグラフ

【図１０】　DLBCL患者の生存時間と診断指標のプロット各種比較の図

【図１１】　実施例２のDLBCL患者の生存時間診断指標のプロット

【図１２】　実施例３の乳癌患者の生存時間診断指標のプロット

【図１３】　実施例３の乳癌患者の変数削除基準としてP≧0.0005を採用したときの生存時間診断指標のプロット

【図１４】　実施例７の乳癌患者の再発時間診断指標のプロット

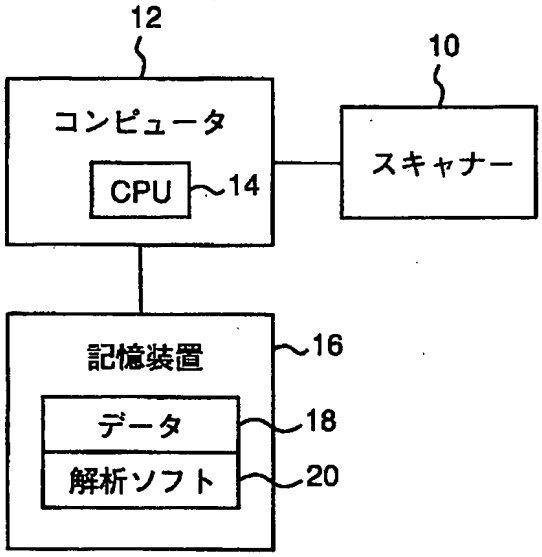
【図１５】　実施例７の乳癌患者の変数削除基準としてP≧0.025を採用したときの再発時間診断指標のプロット

【符号の説明】

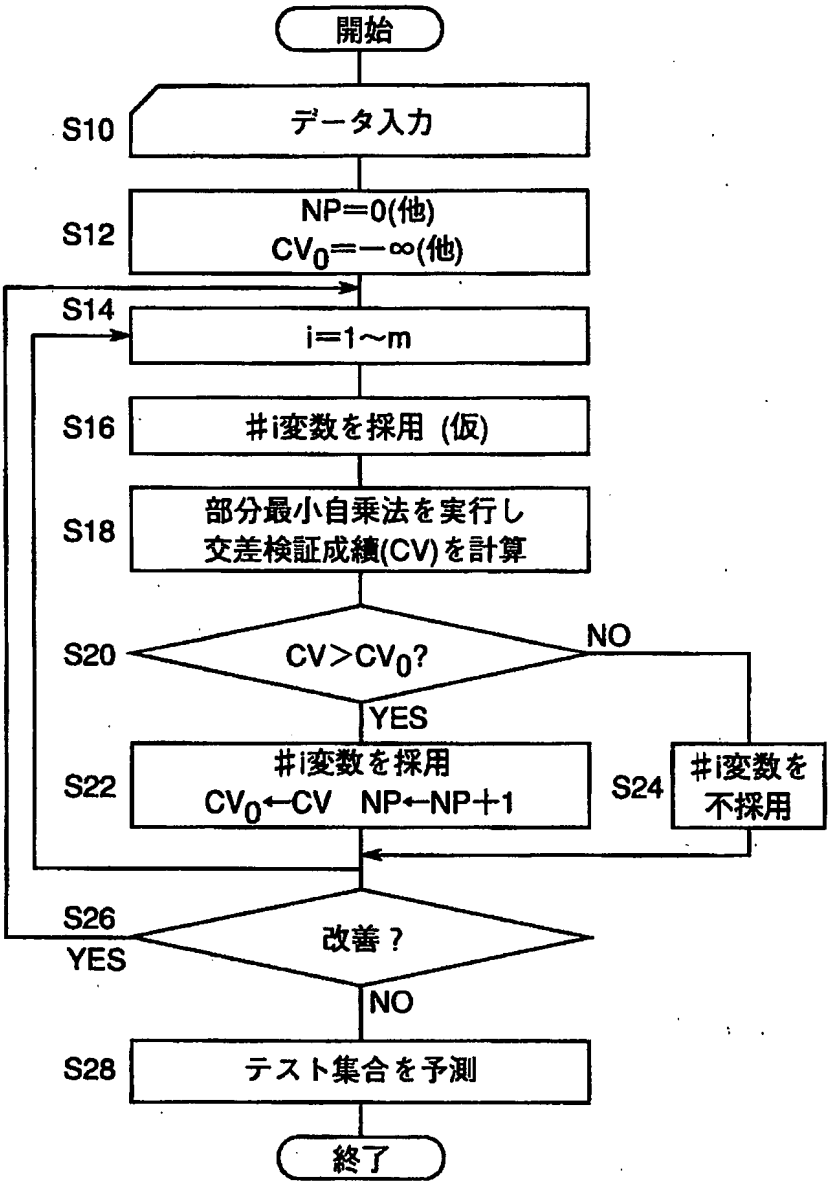
１０　スキャナ、　１２　コンピュータ、　１４　ＣＰＵ、　１６　記憶装置、　１８　記録媒体、　２０　解析ソフト。

【書類名】　図面

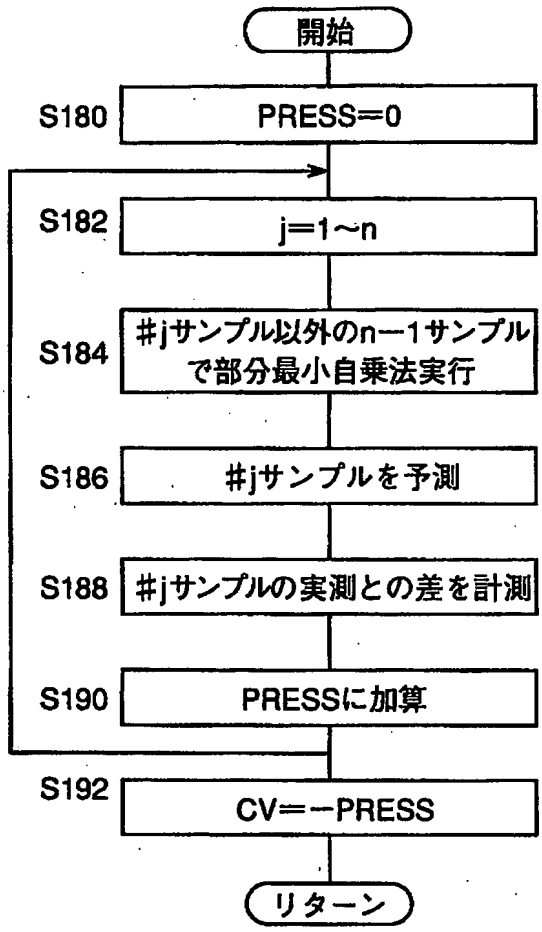
【図１】



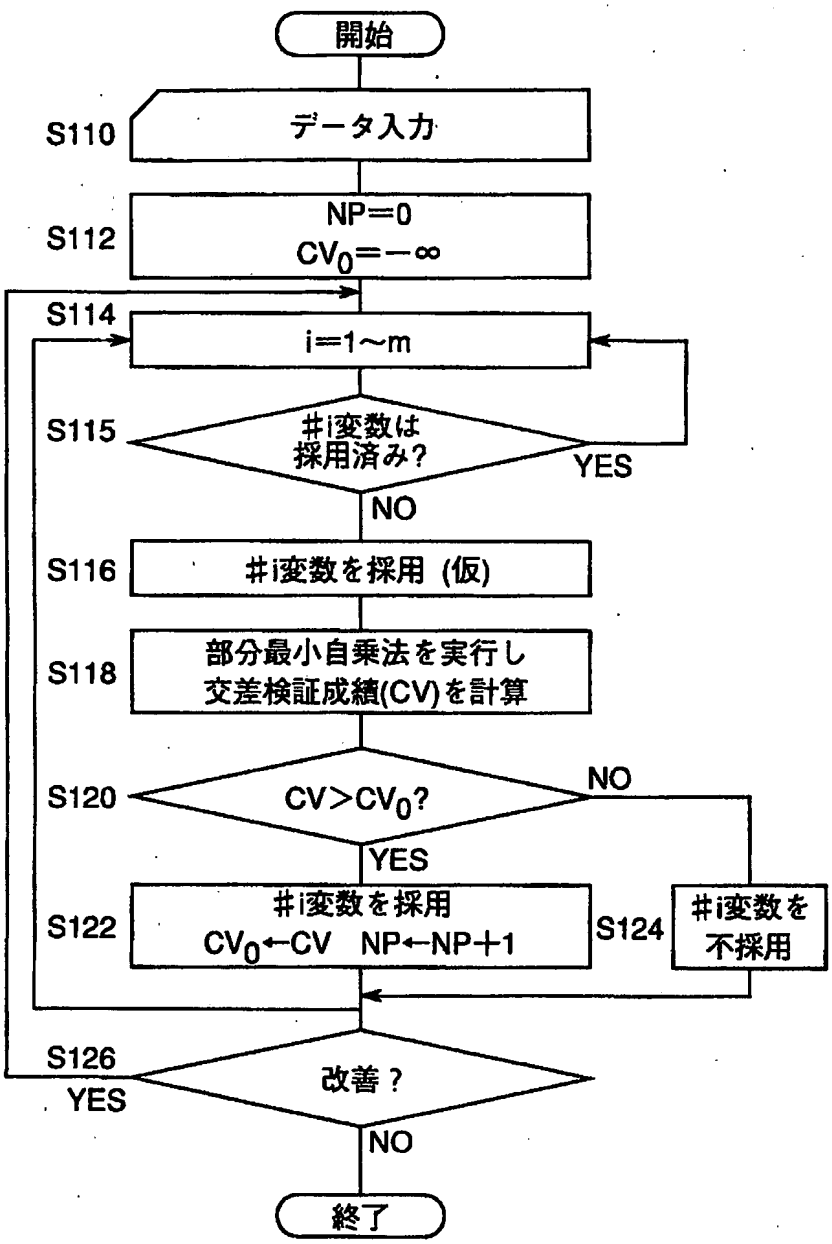
【図２】


開始
S10
データ入力
S12
NP=0(他)
CV0=−∞(他)
S14
i=1～m
S16
#i変数を採用 (仮)
S18
部分最小自乗法を実行し
交差検証成績(CV)を計算
S20
CV>CV0?
NO
YES
S22
#i変数を採用
CV0←CV NP←NP+1
S24
#i変数を
不採用
S26
改善？
YES
NO
S28
テスト集合を予測
終了

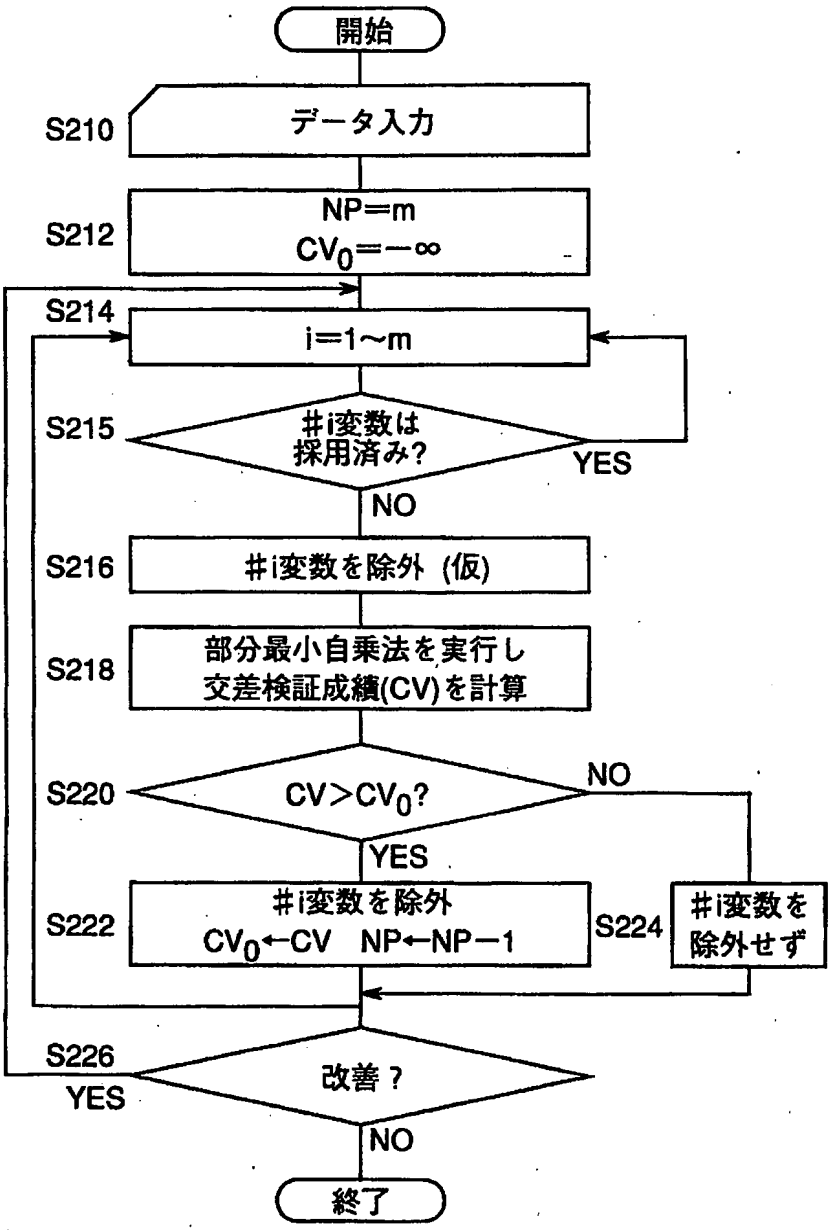
【図３】


開始
S180
PRESS=0
S182
j=1～n
S184
#jサンプル以外のn－1サンプルで部分最小自乗法実行
S186
#jサンプルを予測
S188
#jサンプルの実測との差を計測
S190
PRESSに加算
S192
CV=－PRESS
リターン

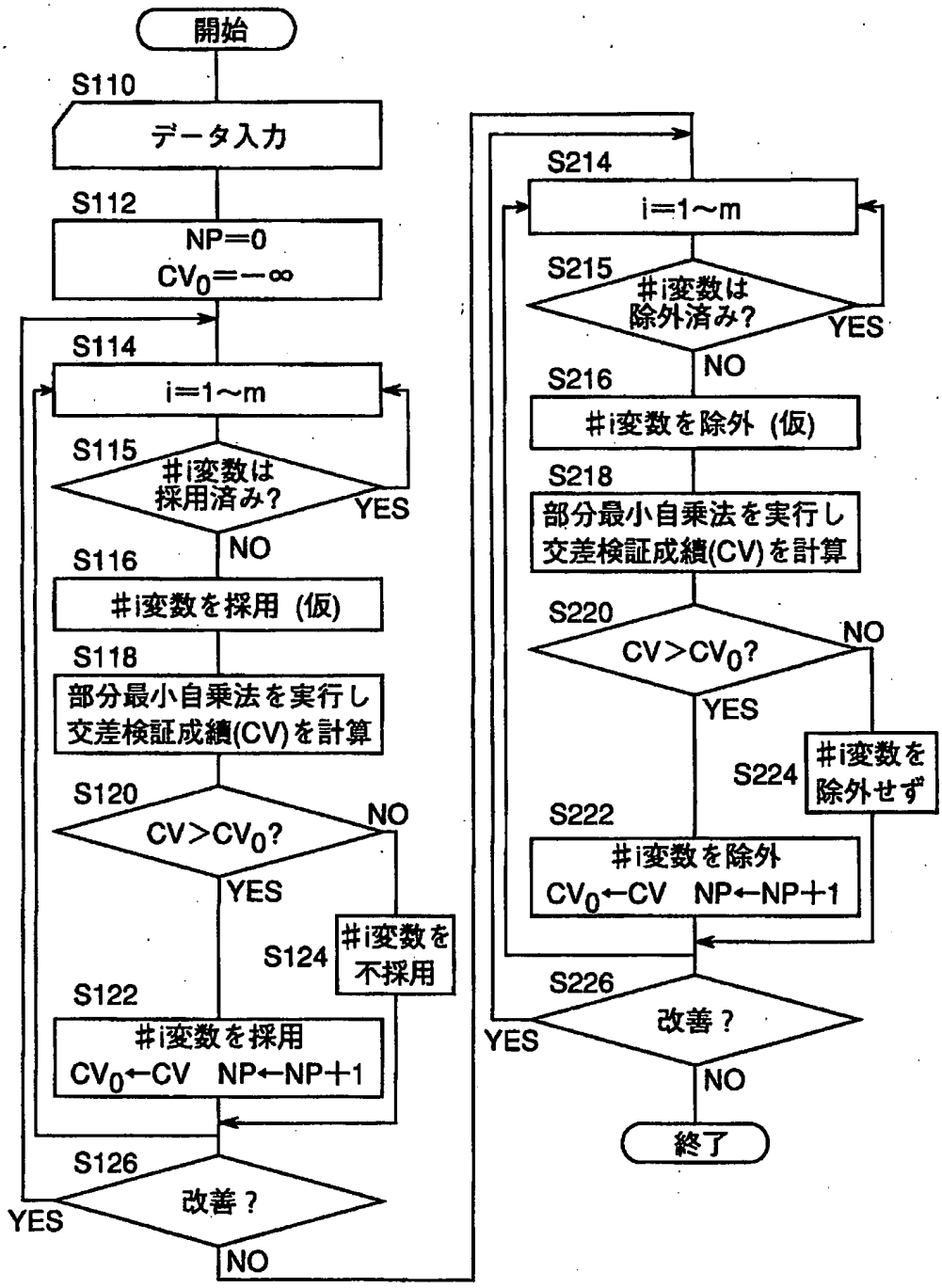
【図４】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1～m
S115
#i変数は
採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し
交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S122
#i変数を採用
CV0←CV NP←NP+1
S124
#i変数を
不採用
S126
改善？
YES
NO
終了

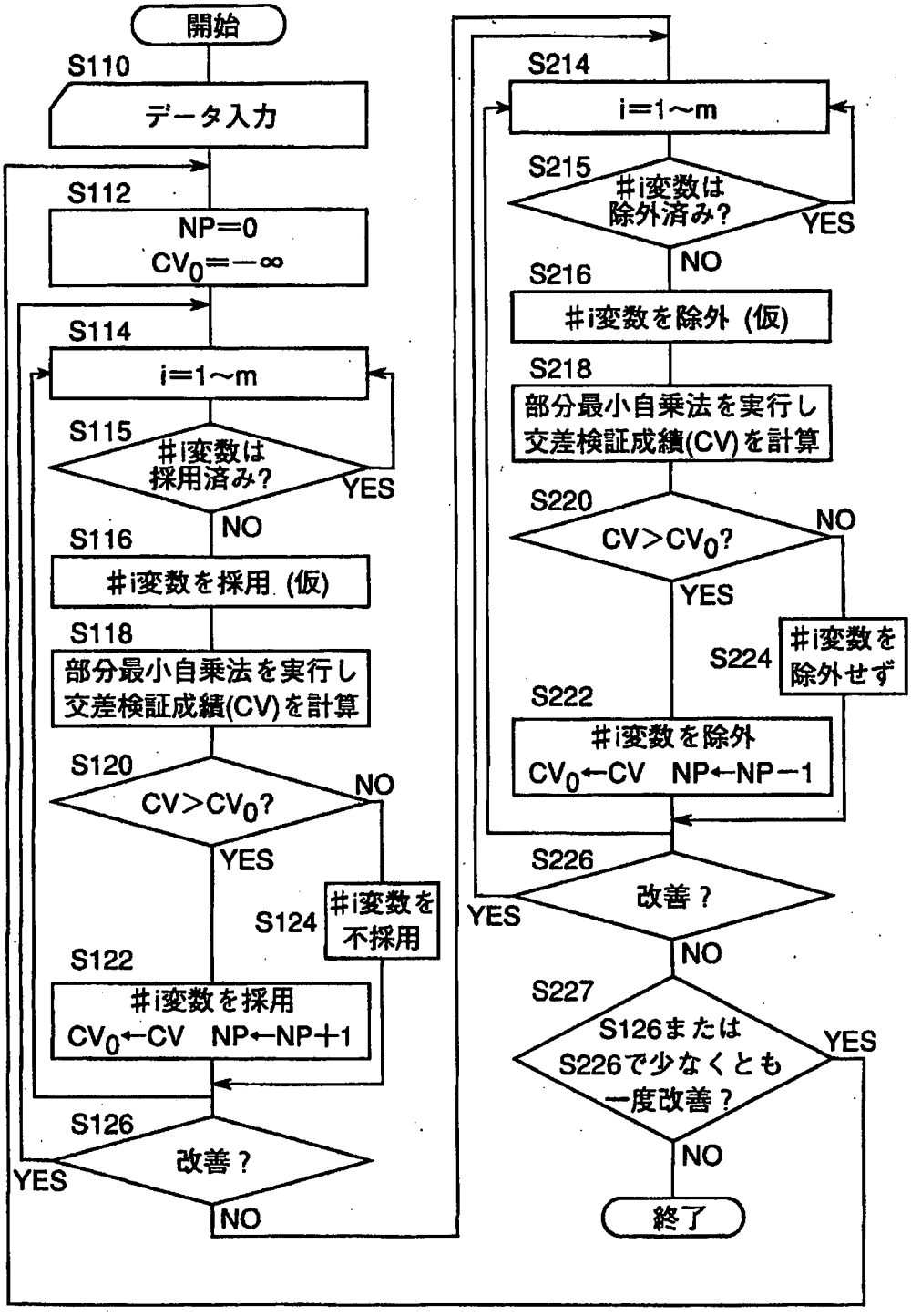
【図５】


開始
S210
データ入力
S212
NP=m
$CV_0=-\infty$
S214
i=1～m
S215
#i変数は
採用済み?
YES
NO
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し
交差検証成績(CV)を計算
S220
$CV>CV_0$?
NO
YES
S222
#i変数を除外
$CV_0 \leftarrow CV$ $NP \leftarrow NP-1$
S224
#i変数を
除外せず
S226
改善？
YES
NO
終了

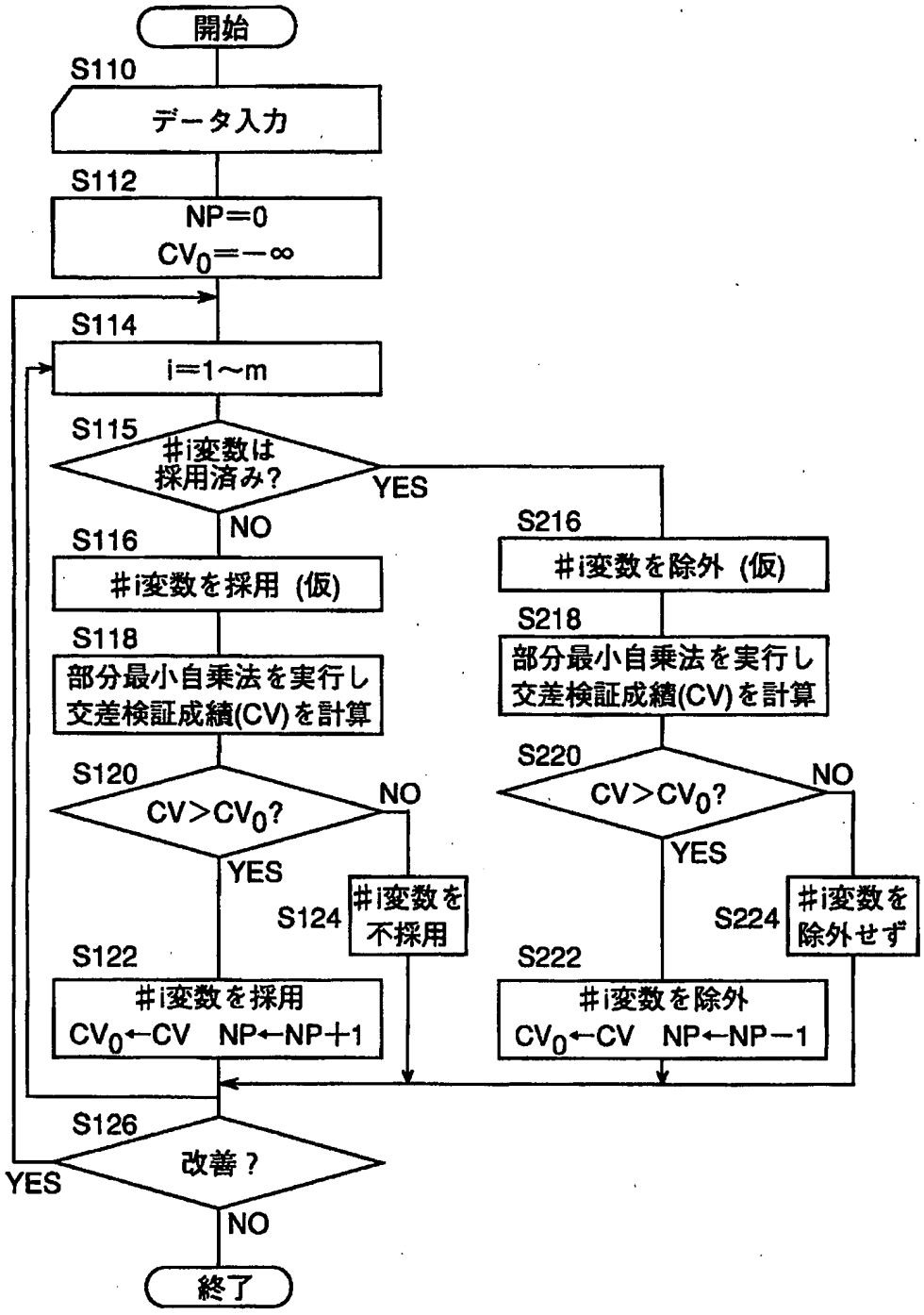
【図６】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1～m
S115
＃i変数は採用済み?
YES
NO
S116
＃i変数を採用 (仮)
S118
部分最小自乗法を実行し交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S124
＃i変数を不採用
S122
＃i変数を採用
CV0←CV NP←NP+1
S126
改善？
YES
NO
S214
i=1～m
S215
＃i変数は除外済み?
YES
NO
S216
＃i変数を除外 (仮)
S218
部分最小自乗法を実行し交差検証成績(CV)を計算
S220
CV>CV0?
NO
YES
S224
＃i変数を除外せず
S222
＃i変数を除外
CV0←CV NP←NP+1
S226
改善？
YES
NO
終了

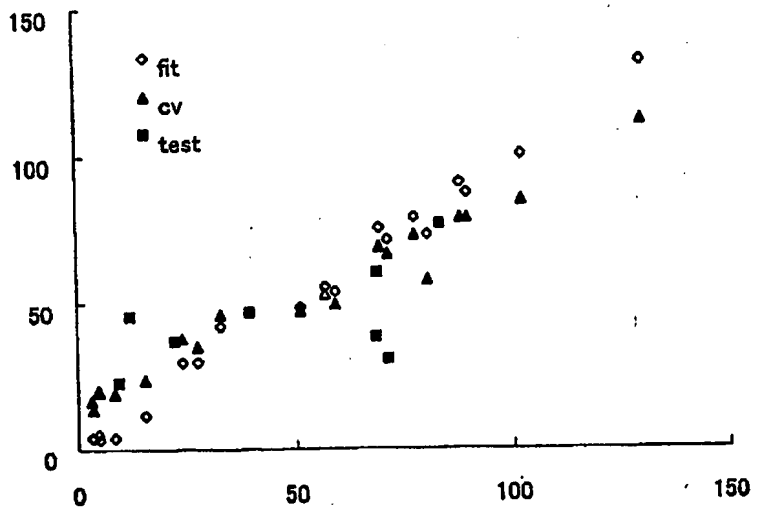
【図７】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1〜m
S115
#i変数は採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S124
#i変数を不採用
S122
#i変数を採用
CV0←CV NP←NP+1
S126
改善?
YES
NO
S214
i=1〜m
S215
#i変数は除外済み?
YES
NO
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し交差検証成績(CV)を計算
S220
CV>CV0?
NO
YES
S224
#i変数を除外せず
S222
#i変数を除外
CV0←CV NP←NP−1
S226
改善?
YES
NO
S227
S126またはS226で少なくとも一度改善?
YES
NO
終了

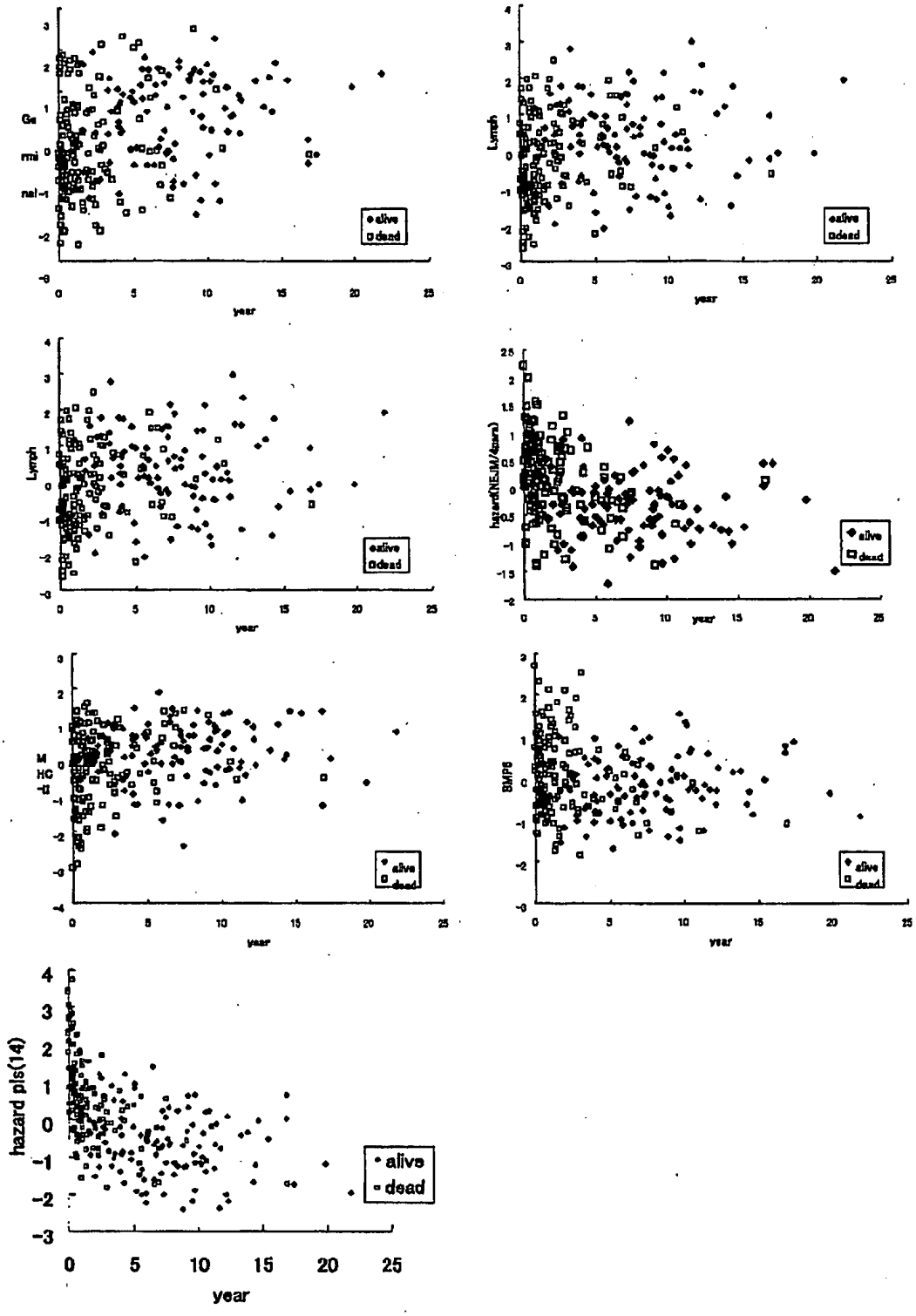
【図８】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1〜m
S115
#i変数は採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S124
#i変数を不採用
S122
#i変数を採用
CV0←CV NP←NP+1
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し交差検証成績(CV)を計算
S220
CV>CV0?
NO
YES
S224
#i変数を除外せず
S222
#i変数を除外
CV0←CV NP←NP−1
S126
改善?
YES
NO
終了

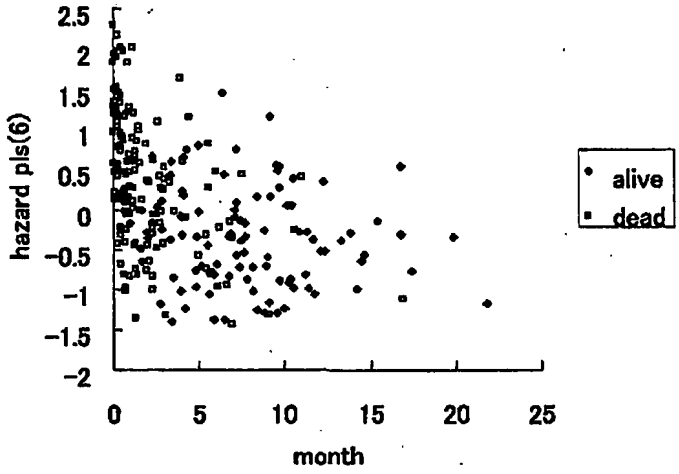
【図９】



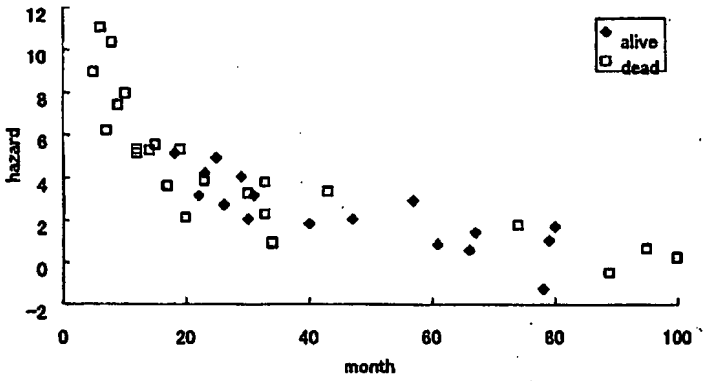
【図１０】



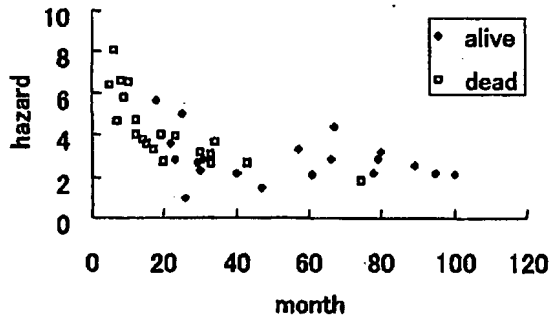
【図１１】



【図１２】

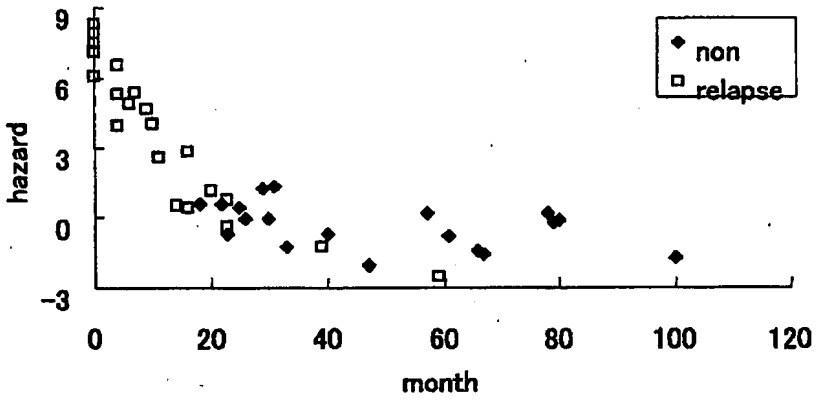


【図１３】

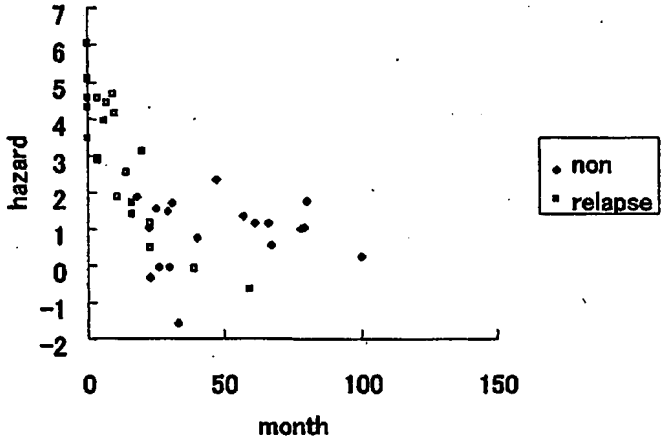


出証特２００３－３０３３９６３

【図１４】


9
6
3
0
-3
hazard
0
20
40
60
80
100
120
month
non
relapse

【図１５】


7
6
5
4
3
2
1
0
-1
-2
hazard
0
50
100
150
month
non
relapse

【書類名】　要約書

【要約】

【課題】　多変量の遺伝子発現情報の効果的な情報処理を提供する。

【解決手段】　生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析において、

生体の状態を目的変数とし、複数(ｍ個)の遺伝子発現の量および／または細胞内物質の量を説明変数とするデータの集合において、データに含まれる説明変数を選択し、選択された説明変数と目的変数とを含む相関モデルについて交差検証成績を計算し、その結果を評価判定する。ここで、交差検証成績が改善しなくなるまで、説明変数の選択、交差検証成績の計算、その結果の評価判定を行い、部分最小自乗法モデルを決定する。

【選択図】　図２

出 願 人 履 歴 情 報

識別番号　〔０００００００３５４〕

１．変更年月日　１９９３年　６月２１日
［変更理由］　住所変更
住　所　大阪府大阪市西区江戸堀一丁目３番１５号
氏　名　石原産業株式会社